Exploring DNA Lesion Recognition Mechanisms for Cyclobutane Pyrimidine Dimers and 6-4 Photoproduct by Rad4

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computational Natural Sciences
by Research

by

Nikhil Jakhar
20171186
nikhil.jakhar@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
(Deemed to be University)
Hyderabad - 500 032, INDIA
September 2023

Copyright © Nikhil Jakhar, 2023
All Rights Reserved

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Exploring DNA Lesion Recognition Mechanisms for Cyclobutane Pyrimidine Dimers and 6-4 Photoproduct by Rad4**" by Nikhil Jakhar, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. Marimuthu Krishnan

To my family, advisor and helpful seniors

Abstract

Genome integrity is seriously threatened by UV radiation's ability to cause damage in DNA. CPD and 6-4PP are two of the most frequent lesions brought on by Ultraviolet radiation. CPD is a DNA lesion occurring from the covalent linking of nearby pyrimidine bases through their C5 and C6 carbons, generating a cyclobutane ring. In contrast, 6-4PP develops when neighbouring pyrimidine bases in DNA become crosslinked following UV exposure, producing a covalent bond between their C6 and C4 carbons. If left untreated, these lesions can result in mutations and even the development of cancer. Clarifying the DNA repair process depends on our ability to comprehend the mechanisms underpinning DNA lesion identification.

DNA damage repair proteins such as XPC/Rad4 recognize and facilitate repair of these damages. In order to better understand how XPC/Rad4 recognises DNA lesions in Nucleotide Excision Repair (NER), this thesis will focus on how it interacts with CPD and 6-4PP. The highly conserved repair process known as NER is in charge of locating and eradicating diverse DNA damages. The current thesis looks at the mechanism, energetics and dynamics of Rad4 sequence-specific lesion recognition. The flipping of a pair of partner bases that are present opposite to the lesion (CPD or 64PP) is one of the crucial biochemical steps that take place while XPC or Rad4 tries to identify and bind to a DNA lesion. Major events in NER like insertion of BHD3-β hairpin, flipping of partner bases and Rad4/XPC association with the damaged DNA were studied by formulating relevant reaction coordinates or Collective Variables. These CVs were then used in Molecular Dynamics and enhanced sampling method, Umbrella Sampling to produce energy profiles for the aforementioned events. A deeper understanding of these energy profiles reveal that insertion of BHD3-β hairpin takes place only after the flipping of partner bases has taken place and a cavity is induced in the active site of the DNA but 6-4PP containing complex promotes the insertion of BHD3-β hairpin with higher ease and flexibility compared to a CPD containing complex.This study also reveals that flipping of 5'-dA partner base takes place before flipping of 3'-dA partner base. And these observations were consistent for both lesions, CPD and 64PP.

The results of this study will be useful to improve our knowledge of the molecular processes behind Rad4's identification of DNA lesions and give light on the critical steps in the NER pathway. Finally, this information may aid in the creation of focused therapeutic approaches for preventing and treating UV-induced DNA damage, enhancing our capacity to protect genomic integrity and lessen the dangers of UV radiation exposure.

Acknowledgements

I would like to convey my sincere gratitude to the people listed below who have contributed significantly to my academic career.

Above all else, I want to express my sincere gratitude to my advisor Prof. Marimuthu Krishnan, whose direction, knowledge, and constant support have greatly influenced my research and general academic development. His commitment to my success and dedication have been a true inspiration to me.

I also want to express my sincere gratitude to my parents for their unwavering support, unending sacrifices, and everlasting love for me. They have been a constant source of inspiration for me, and I will always be grateful for their unflinching support.

To my encouraging sister, Kanika Jakhar i would like to express my gratitude for always being there to listen, offer words of support, and share insightful observations. Having you by my side has given me courage during trying moments because of your unfailing faith in my talents.

Additionally, I would like to acknowledge the valuable contributions and guidance provided by my seniors, co authors Akshay Prabhakant and Hemanth Vemuri and faculties of IIIT Hyderabad. Their expertise, experience, and willingness to impart it on me have been priceless resources for me as I've struggled to overcome a variety of scholastic and personal obstacles. I am also grateful for my labmate Keshavan Seshadri's constant support and ideas.

And finally, i would like to thank IIIT Hyderabad's HPC centre and particularly Dr. Semparithi Aravindan for their great assistance. Their kind offering of computer resources made it possible for me to complete my thesis.

I sincerely appreciate everything these individuals and organisations have performed for me academically.

Contents

List of Figures

Figure Page

Chapter 1

Introduction

Biological systems come in all shapes and sizes, from simple single-celled organisms to intricate multicellular ones. Notwithstanding their differences, all living things are subject to the same basic physical principles. Thermodynamics, kinetics, and mechanics are only a few of the fundamental theories that regulate biological processes at the molecular level. To fully understand the molecular mechanisms underlying many biological activities, it is crucial to comprehend how these principles materialise in living systems [1].

The intricate process of cellular self-organisation at the molecular and physical levels is important for the proper functioning of biological systems. Researchers can develop new cures and treatments for a variety of diseases and disorders by understanding the mechanisms and processes that control cellular organisation. This is done through gaining insights into a wide range of biological processes [2]. Cells, that are the basic building blocks of life, are composed primarily of two types of macromolecules: proteins and nucleic acids [3].

A significant area of biology research is the study of proteins in cells, which has profound effects on our comprehension of biological functions. They are essential components of biological cells, and are involved in several cellular processes, from signal transmission between cells to structural support and chemical reaction catalysis. In addition to taking part in different signalling pathways, they carry other chemicals throughout the cell [4].

Ribonucleic acid (RNA) and deoxyribonucleic acid (DNA) are the two primary types of nucleic acids, which are a crucial part of all cells. DNA is mostly found in prokaryotic cells' cytoplasm and the nucleus of eukaryotic cells. It acts as a genetic information carrier and holds the instructions required for protein synthesis. RNA, which is created from DNA, performs important roles like gene regulation and controlling protein synthesis [5]. DNA generates proteins through a process known as protein synthesis, or translation. The processes that make up this process include transcription, RNA processing, translation initiation, elongation, and termination. The entire procedure is carefully regulated and includes a number of chemicals and protein building blocks [6].

Gene expression, DNA replication, DNA repair, and chromatin remodelling are only a few biological processes that depend on interactions between the aforementioned proteins and DNA. Many DNA-binding proteins control these relationships by recognising particular DNA sequences and interacting with them in a sequence-specific way. DNA-protein interactions can change the structure and function of the DNA molecule, and these alterations can have a significant impact on gene expression and other cellular processes [4, 7, 8]. Thus, it is essential to comprehend these interactions in order to comprehend the basic functions of life.

Protein-DNA interactions have been studied using a variety of experimental approaches, each with different advantages and drawbacks. X-ray crystallography, NMR spectroscopy, fluorescence resonance energy transfer, electrophoretic mobility shift assays, and surface plasmon resonance are some of the techniques used to study these interactions. However, due to the complexity of the interactions, the dynamic nature of the complexes, and the difficulty in isolating and characterising individual interactions, none of them provide a detailed image of protein-DNA interactions. They allow for a level of data that exceeds experimental timelines in the investigation of conformational changes that take place in both proteins and DNA upon complex formation. They allow for a level of data that exceeds experimental timescales in the investigation of conformational changes that take place in both proteins and DNA upon complex formation. [9, 10].

This study utilizes modern computational techniques including molecular dynamics and accelerated sampling methods to explore protein-DNA interactions at the molecular scale. It explores molecular mechanism, dynamics and energetics behind DNA damage recognition process by proteins.

1.1 Proteins

Proteins, an essential class of biomolecules, play a crucial role in how living things function. Many proteins serve as enzymes that accelerate cellular chemical processes [4, 7, 11]. Numerous biological functions like protein synthesis, DNA replication, and metabolism, use these enzymes. In the well developed cells like cells in humans, proteins make up the cytoskeleton, which gives the cell its shape and helps it maintain its structural integrity. They construct a number of structures inside the cell, including the ribosomes, the nuclear envelope, and numerous organelles. Also, they are involved in the transport of molecules within the cell, as well as across cell membranes [12]. Its further roles include serving as signalling molecules and communicating between and within cells [4].

All the proteins that an organism expresses are included in the proteome. The proteome of humans, for example, contains about 32,000 distinct proteins, as opposed to, say, the yeast Saccharomyces cerevisiae, whose proteome contains only 6,000 proteins [13]. Since it is a challenging and complex method that necessitates knowledge in many different fields like molecular biology, biochemistry, crystallography, and computational biology, only a tiny percentage of proteins have the whole 3D structure identified experimentally. Even when direct experimental data of a protein is not available, homology-based techniques can be employed in research on protein function and structure [14]. Examining a protein's interactions with other proteins can help us better understand its role in unidentified, unclassified proteins. Because interacting proteins frequently have similar functions, discovering the proteins that interact with an unknown protein can reveal information about its function [15]. If it is found that an unknown protein interacts with proteins involved in DNA repair, for example, it may play a role in the DNA damage response.

The primary building blocks of proteins are linear chains made of 20 distinct naturally occurring amino acids [4]. Despite having minimal building blocks, proteins are capable of performing a staggering array of different jobs. A protein's 3-D structure, which is influenced by its amino acid sequence and interactions with its environment, determines its function. It's 3-D structure is formed by the folding of the linear chains mentioned above.

The basic units of proteins are amino acids. A core carbon atom, an amino group (-NH2), a carboxyl group (-COOH), and a variable side chain or functional group (-R) that distinguishes each amino acid are the components of their shared structure. There are a total of 20 frequently occurring amino acids in proteins, and each one is distinguished by its particular chemical structure or functional group. Amino acids link together to make up a linear chain known as a polypeptide by peptide bonds. Peptide bonds are developed between an amino acid's amino (-NH2) group and its carboxyl (-COOH) group in another amino acid. As a result, a polypeptide chain is created, which has a carboxyl terminus ,C-terminus, at one end and an amino terminus , also known as N-terminus, at the other. There are 4 levels of protein structure that outlines the organization of a protein from its primary sequence of amino acids to its three-dimensional shape. A protein's linear sequence, which is determined by the configuration of peptide bonds, is referred to as the primary structure of a protein. Peptides are the name given to smaller chains of amino acids, usually containing 20 to 30 amino acids, whereas polypeptides are the name given to longer chains.

The polypeptide chains mentioned above twist and fold to form the secondary structure of protein. Hydrogen bonding or H-bonds between the carbonyl and amino groups of two amino acids stabalize the secondary structure. The development of this structure also involves van der Waals and electrostatic interactions between the residues (-R). The two most common types of secondary structure are α helices and β sheets. As shown in Figure 1.2, α helices are tightly coiled, while β sheets are made up of strands that are folded back and forth on themselves. The protein backbone's remaining segments are formed by small coils and turns.

α helices are right-handed coiled structures that resemble a spring or a spiral staircase (see Figure 1.2). The carbonyl group of one amino acid residue and the amino group of another amino acid residue that is four residues further along the chain establish hydrogen bonds in order to increase stability. The amino acids' side chains, often known as their -R groups, project outward from the helix

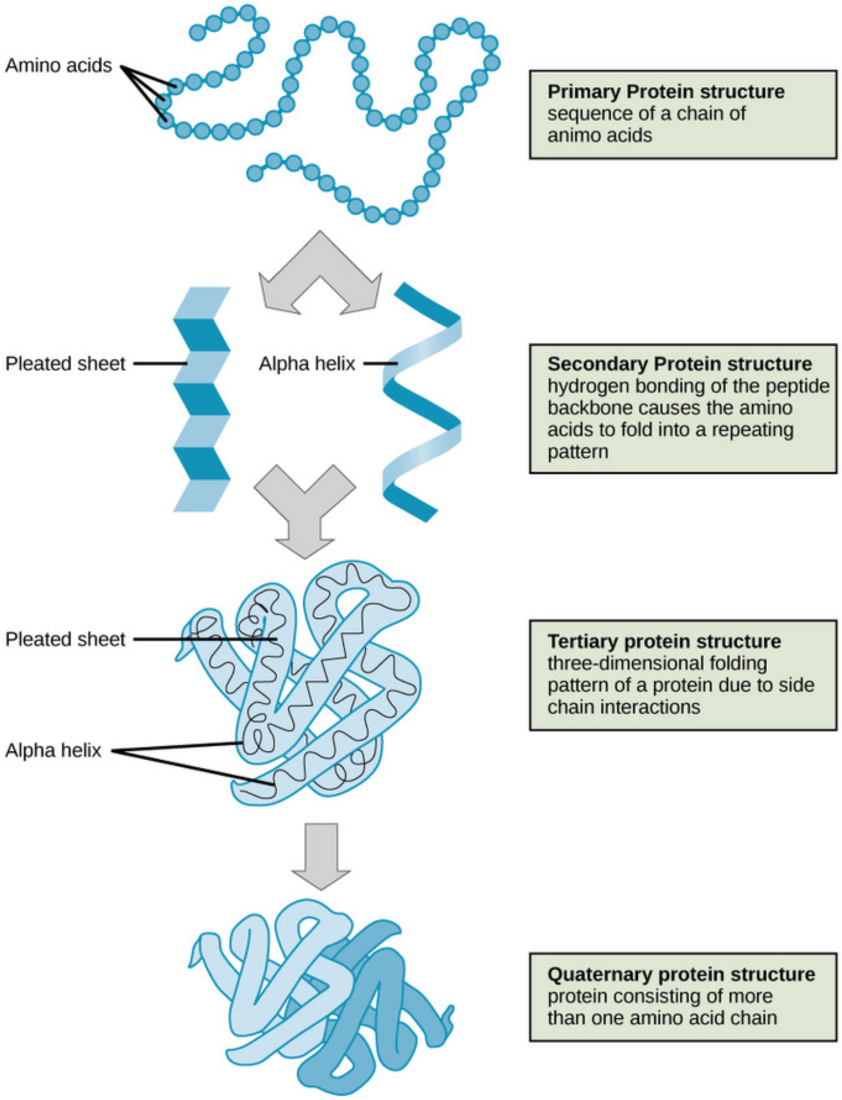

Amino acids
Primary Protein structure
sequence of a chain of animo acids
Pleated sheet
Alpha helix
Secondary Protein structure
hydrogen bonding of the peptide backbone causes the amino acids to fold into a repeating pattern
Pleated sheet
Alpha helix
Tertiary protein structure
three-dimensional folding pattern of a protein due to side chain interactions
Quaternary protein structure
protein consisting of more than one amino acid chain

Figure 1.1: Protein structure as well as function are determined by the four degrees of protein complexity. The linear polypeptide chain formed by the arrangement of amino acids is the fundamental or primary structure. The secondary structure of a polypeptide chain is the consistent, repeated configuration of adjacent amino acids which forms α helix or pleated sheats. When all the components of the secondary structure have folded into one another, the polypeptide chain's overall three-dimensional shape is known as the tertiary structure. The positioning and configuration of a protein's subunits with respect to one another is referred to as its quaternary structure. Illustration and text adapted from: `https://www.coursehero.com/study-guides/boundless-chemistry/protein-structure/`.

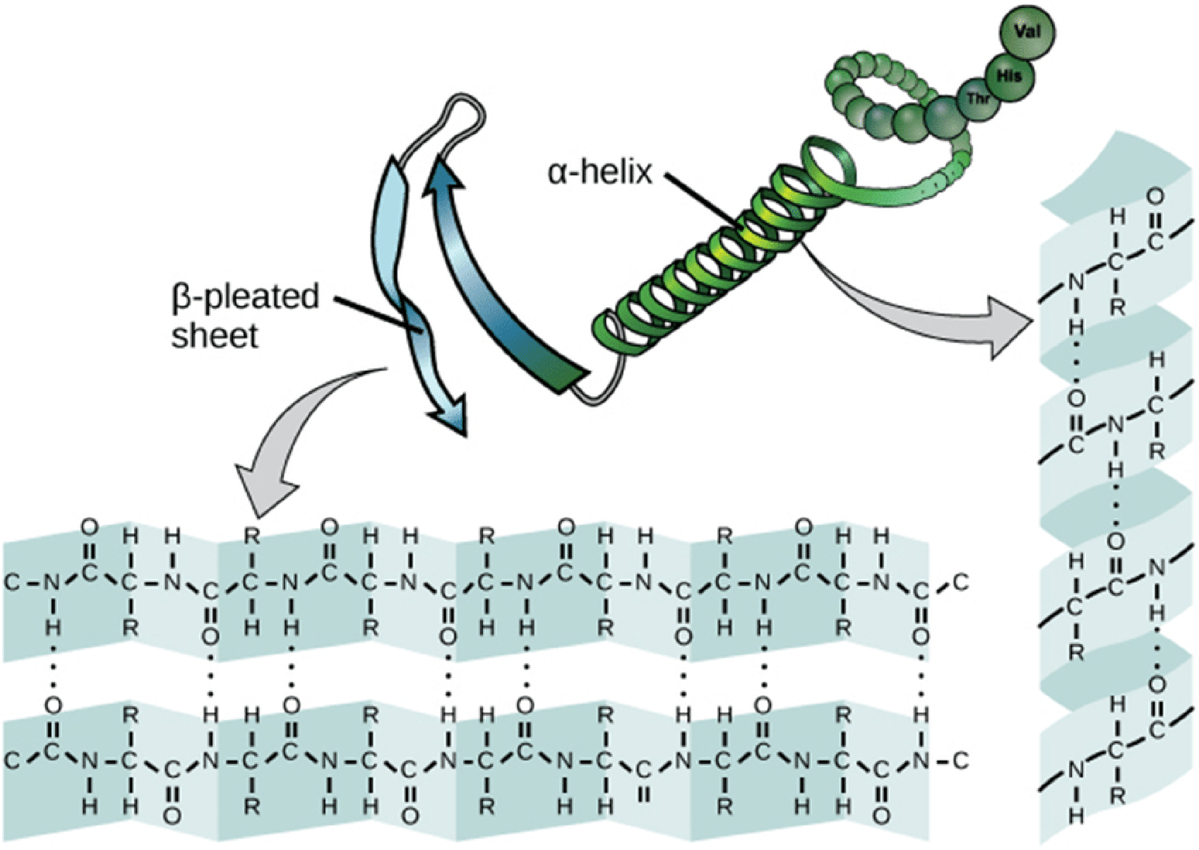


Figure 1.2: Secondary structure units of proteins, α-helices and β-sheets. Illustration and text adapted from [16]

structure. β sheets, on the other hand, consist of a series of β strands, which are extended, zigzagged strands that can be arranged in parallel or antiparallel orientations. The carbonyl group (-COOH) of one strand and the amino group (-NH2) of a nearby strand form hydrogen bonds, which join the β strands together. The resulting β sheet is pleated, with the strands running perpendicular to the sheet's axis. β sheets can either be parallel, in which case the strands go in the same direction, or antiparallel, in which case they do the opposite. The hydrogen connections between the strands stabilise the sheet and provide a planar structure in both circumstances.

The fully folded polypeptide chain's (including any α helices, β sheets, and other secondary structures, as well as any additional loops or folds that may be present) 3D shape is known as a protein's tertiary structure. Numerous interactions between the side chains of the amino acids, such as hydrogen bonds between polar ones and hydrophobic interactions between non-polar ones, help to stabilise the tertiary structure [4].

It was discovered through tertiary structural analysis of numerous proteins that specific secondary structure combinations are shared by many proteins. These motifs or folds are typical characteristics that frequently affect a protein's chemical function, especially for small proteins where the number of secondary structure combinations is relatively constrained. For larger proteins, the picture is more complex, as there are many more possible combinations of secondary structures. For them these folds or motifs are compactly folded and constitute the so called "domains". Domains are discrete regions within a protein that fold independently and have a unique structure, sequence, and function. They are preserved among proteins from the same functional or structural families. Domains carry out particular tasks such binding to other proteins, DNA or RNA, initiating chemical events, or controlling the activity of proteins. Experimental techniques like X-ray crystallography or NMR spectroscopy can be employed to identify and characterize them.

The arrangement of many polypeptide chains into a useful protein complex is known as the quaternary structure of a protein. The quaternary structure is supported by a variety of interactions, including hydrogen bonds and hydrophobic interactions, much like the tertiary structure is. By coordinating with other molecules, this globular structure is able to carry out intricate, multi-stage activities. Function of proteins with many subunits, such as enzymes, antibodies, and haemoglobin, depends on the quaternary structure.

In essence, proteins are intricate and adaptable macromolecules that are essential for nearly every element of life, from cellular functions to organismal function. They provide a rich area of study for academics in many different domains due to their diversity in structure and function.

1.2 Nucleic Acids

Deoxyribonucleic acid also referred to as DNA, is a biological molecule that houses the vital genetic information required for the development and survival of all living things. It is made up of smaller units known as nucleotides and has the structure of a long, double-stranded helix. The 3 components of each nucleotide are a sugar molecule, a phosphate group, and a nitrogenous base. There are 4 nitrogenous bases: adenine (A), thymine (T), guanine (G) and cytosine (C). The aforementioned bases pair up with each other through hydrogen bonding, with A always bonding with T, and G always bonding with C, to form the "rings" of the DNA ladder [17].

In 1944, Avery, McCarty, and MacLeod provided strong evidence that DNA is the genetic material [18]. Their discoveries were important as they paved the door for more exploration into the structure and function of DNA and helped to confirm DNA as the molecule in charge of transmitting genetic information. A few years later, in 1950, Chargaff observed that while the proportions of the 4 nitrogenous bases in DNA varied between different species, they were always roughly proportional. Specifically, Chargaff found that the amount of adenine in DNA always equal to the amount of thymine, and the amount of guanine is always equal to the amount of cytosine (C) [19]. It was important because it made it possible for Francis Crick and James Watson to postulate the double helix structure of DNA in 1953 [20]. The double helix's shape implied that the two DNA strands were complementary, with the nucleotides on one strand dictating the bases on the other (see Figure 1.3). This suggested that the bases may be read like a code, with each "word" (a codon) being made up of three bases and designating a certain amino acid in a protein. In 1961, Nirenberg and Matthaei were able to determine the specific codons that corresponded to each of the 20 amino acids found in proteins [21].

The Watson-Crick model has four following key characteristics, which have undergone numerous modifications and improvements since the original:

1. According to the Watson-Crick model, DNA has two antiparallel strands that are twisted into a helical form, making it a double-stranded molecule. Based on information gleaned from X-ray crystallography, which revealed that DNA has a regular, repeating pattern that suggested a helical structure, this structure was inferred. For DNA replication and the transmission of genetic information, complementary base pairing is made possible by the double-stranded structure.

2. The fact that DNA is a right-handed helix causes it to appear to be twisting clockwise when viewed from one end. A left-handed helix, on the other hand, would twist the opposite way, clockwise. DNA's right-handed helix structure is significant because it enables the cell to pack the DNA molecule tightly.

3. DNA has two strands that are antiparallel to one another, which means they move in different directions. Due to this, one end of each strand has a free 3' hydroxyl group, whereas the other end has a free 5' phosphate group. The 5' end of one strand is coupled with the 3' end of its complementary strand in the complementary base pairs (A-T and C-G) (as shown in Figure 1.3). This configuration is crucial because it enables correct DNA replication during cell division.

4. Interstrand H-bonds between the complementary base pairs (A-T and G-C) and stacking interactions between the nucleobases (see Figure 1.3) throughout the helix work together to keep the DNA double helix stable. Between the nitrogenous bases on opposing strands, hydrogen bonds form, and between the flat surfaces of the nucleobases, stacking interactions take place [22–24]. The DNA molecule's ability to carry and transfer genetic information depends on these interactions both for preserving the molecule's structure and for doing so.

The entire set of genetic instructions contained in a cell's DNA is known as the genome. It includes all the details required to create and sustain an organism, including its morphological and functional characteristics [25]. The genome is located in the cell nucleus of eukaryotic cells, which is encircled by a double membrane structure known as the nuclear envelope. The confinement of DNA within the nucleus of eukaryotic cells offers a protective, structured, and controlled environment that aids in ensuring the integrity and appropriate operation of the genetic material as nucleus provides a physical

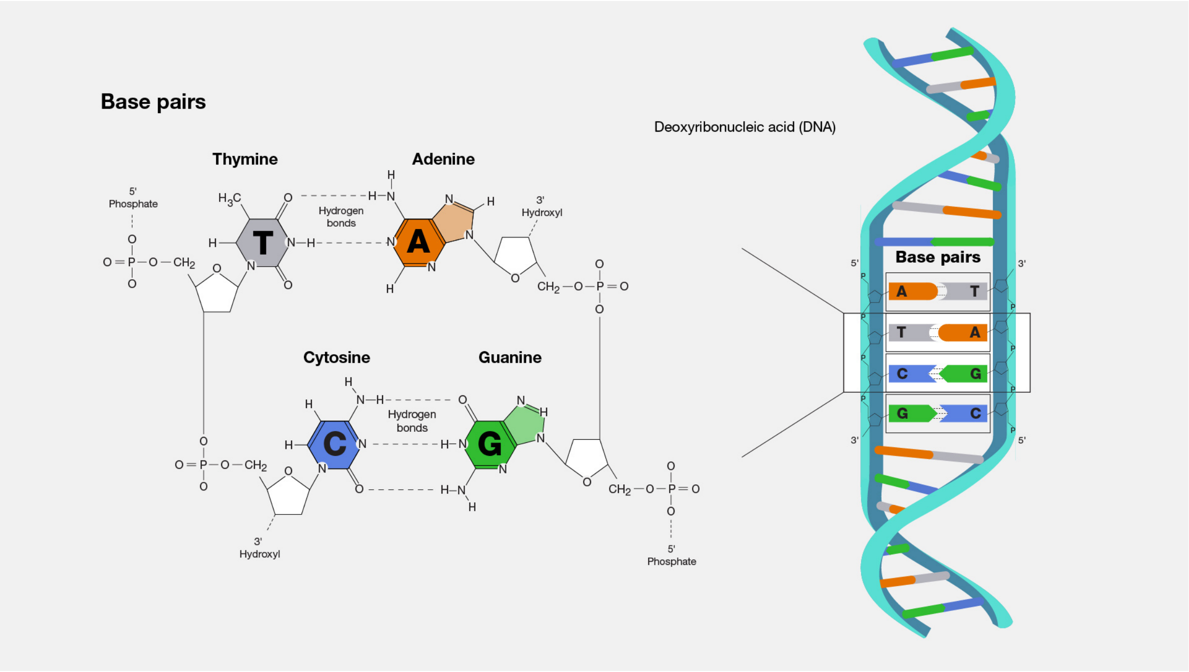


Figure 1.3: Base pairing in DNA. Three H bonds link guanine to cytosine, while 2 H-bonds link thymine to adenine. Deoxyribose and phosphate groups alternately form the backbone of each strand and aforementioned bases are attched to these sugar molecules. Illustration and description adapted from `https://www.genome.gov/genetics-glossary/Base-Pair`.

space for the chromatin (DNA and associated proteins) to be tightly packaged and regulated, allowing for precise control of gene expression inside a selective barrier.

There are around 3 billion base-paired nucleotides in a regular human cell. These base-paired nucleotides are responsible for encoding the information necessary for the synthesis of roughly 30,000 distinct proteins, which play a part in numerous cellular processes and functions [25]. For the tremendous quantity of genetic information to squeeze into the tiny space within a cell, the compact packing of this enormous amount of DNA is necessary and nucleus's ability to compactly house the genome depends on DNA's elasticity and flexibility.

The bending preferences of DNA can influence the three-dimensional structure of the DNA molecule and its interactions with other molecules. It can also influence the binding of architectural proteins, such as histones [26]. The histone proteins and the DNA molecule combine to form a structure known as a nucleosome. The synthesis and stability of the nucleosome are critically dependent on the DNA molecule's ability to bend.

The flexibility and elasticity of the DNA molecule are essential for allowing the transcription and replication machinery to move along the DNA strand as in both these processes the double helix unwinds and separates to create a bubble-like structure. This allows the RNA polymerase enzyme to access and copy the DNA sequence for a specific gene. Moreover, specialised proteins that are known as transcription factors attach to particular DNA regions to control the expression of genes nearby [27, 28]. The precise regulation of gene expression depends on these proteins' capacity to interact with the DNA sequences they need to bind to, which is made possible by the DNA molecule's ability to bend and twist.

Proteins that bind to DNA typically have specific domains or regions that recognize and bind to particular sequences or structural features in the DNA molecule. Yet, in the vast and largely fluid nuclear environment, proteins must first find and detect the target location on the lengthy DNA strands before binding can occur [29]. Proteins include both local DNA sequences and other global substructures as part of their recognition mechanisms. The two major classes of recognition mechanisms used by proteins are:

1. **Direct readout**: By direct interactions between the protein's amino acid residues and the nucleotide bases, the protein in this process detects a particular nucleotide sequence in the DNA. Direct readout is often used by proteins that bind to the major groove (see Figure 1.4) of DNA [29–32]. In majority of cases, direct readout of DNA involves the protein directly contacting the DNA molecule and inserting a DNA-binding motif, such as an α helix or β hairpin, into the major or minor groove of the DNA helix. Varying exocyclic groups of the paired-up bases are accessible in the major and minor grooves of DNA, which results in various charge patterns and that in turn helps the motifs to differntiate between T-A and A-T base pairs or G-C and C-G base pairs in a sequence [29, 33–35].

2. **Indirect readout**: In this mechanism, the protein recognizes the structure or conformation of the DNA molecule rather than its sequence. Indirect readout can also occur through the recognition of DNA shape or deformations, such as bends or kinks in the DNA helix. Here, a chain of bases not in touch with the protein controls the stability and specificity of the way the enzyme may attach to its target DNA [29–32]. Indirect readout mechanism that enables a limited set of damage repair proteins to identify structural anomalies in the DNA helps them recognise a variety of DNA damages.

In order to achieve precise DNA binding, certain proteins combine direct and indirect mechanisms of recognition. Proteins can differntiate target sequences from non-target sequences with a high degree of affinity and selectivity, thanks to the combination of these recognition techniques.

Due to the hydrophilic property of the DNA double helix, a layer of water molecules is always present on its surface. With the exposed phosphate and hydroxyl groups of the DNA backbone as well as the bases in the major and minor grooves, these water molecules create hydrogen bonds [37, 38]. The strength of these hydrogen bonds varies with the changing base pairs and the chemical environment. The water molecules around the DNA molecule can affect the affinity with which proteins bind to DNA as they compete against one another for hydrogen bonding sites on the surface of the double helix. So

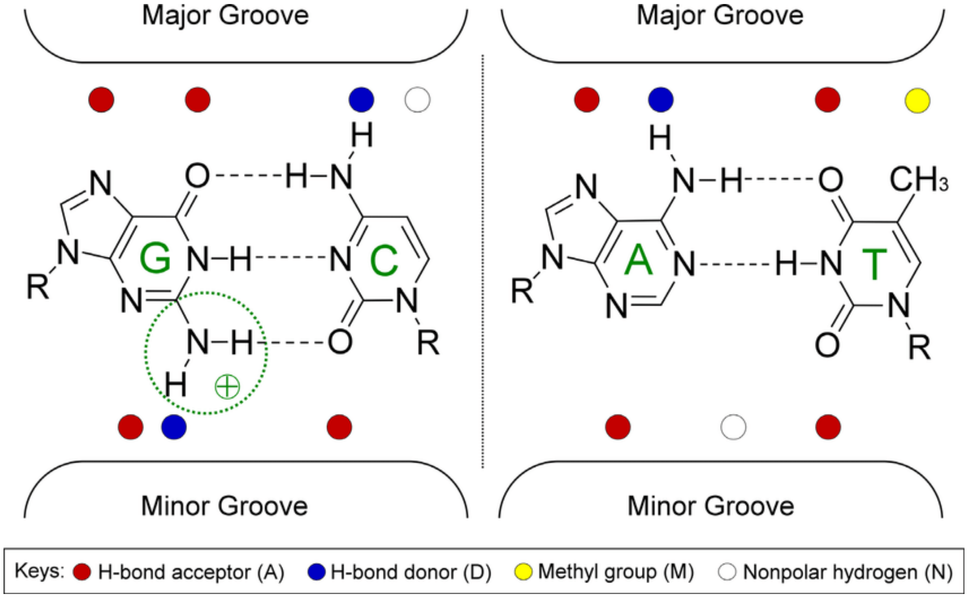

Major Groove
Major Groove
G
C
A
T
CH3
R
Minor Groove
Minor Groove
Keys: ● H-bond acceptor (A) ● H-bond donor (D) ● Methyl group (M) ○ Nonpolar hydrogen (N)

Figure 1.4: C/G and A/T base pair functional groups are visible on the main and minor grooves. In the major groove (direct readout), interactions with unique functional groups are the primary mechanism by which proteins recognise binding sites, nonetheless, the minor groove's pattern deviates from the norm. Illustration and caption adapted from [36]

in simple terms water molecules create a barrier that must be overcome by proteins that wish to bind with the DNA. This can have an impact on the energetics of the proteins binding to DNA, which makes the dynamics of water molecules around the DNA an area of interest when studying the interactions between protein and DNA.

1.3 Protein-DNA Complexes

A biomolecule's biological function is significantly influenced by its macromolecular structure. For a protein or nucleic acid to interact with other molecules in the cell and perform its unique function, its complex 3-D structure is crucial.

When proteins interact with DNA, they can control a variety of biological processes. Some examples of processes that can be regulated by protein-DNA interactions include gene expression, DNA replication and DNA repair [39]. While some structural proteins can bind to any DNA sequence, the majority of other proteins that come in contact with DNA bind to specific DNA sequences [40].

Some proteins only have one DNA-binding domain (DNA-binding domains are regions of the protein that are specialized to recognize and bind to DNA), whereas others have multiple domains that recognise different DNA sequences or interact with different parts of the same DNA molecule. Because of variations in base composition and sequence, different DNA sequences can adopt different structures and might not always exist in the Watson-Crick double helix structure. Proteins can recognize these structural variations that alter the surrounding electrostatic potential as the minor groove of B-DNA is narrow and deep, while the major groove is wide and shallow. The electrostatic potential of the grooves varies depending on the form and size of these grooves. As mentioned in the earlier sections proteins can recognize specific DNA sequences mainly using 2 types of mechanisms, direct readout and indirect readout. In indirect readout, by interacting with certain structural elements in the DNA molecule, such as grooves, bends, or kinks, which might differ depending on the nucleotide sequence, proteins can detect DNA sequences. In direct readout, they recognize specific DNA sequences by directly interacting with the chemical groups on the nucleotide bases.

Hydrophobic, electrostatic, and hydrogen bonding interactions frequently work together to stabilise protein-DNA interactions. Positively charged amino acid residues like lysine and arginine interact with the negatively charged phosphate backbone of DNA through electrostatic attraction. Furthermore, hydrogen bonds can be formed between particular protein amino acid residues and DNA nucleotide bases. For instance, the nitrogenous bases (adenine, thymine, cytosine, and guanine) and particular amino acids such asparagine, glutamine, and histidine can form hydrogen bonds [41, 42]. Other than electrostatic and hydrogen bonds hydrophobic interactions are also present in protein DNA complexes. Leucine, isoleucine, and other hydrophobic amino acids, including phenylalanine, can interact with the hydrophobic regions of DNA. When nonpolar amino acids group together to reduce their exposure to nearby water molecules, these interactions take place [32]. The protein-DNA complex is more stable as a result of these interactions.

In addition to the aforementioned interactions, the shape of protein and DNA plays an important role in stabalising protein-DNA complexes. Proteins that interact with DNA often possess structural motifs, such as helix-turn-helix or zinc fingers, that enable them to identify and bind to particular DNA sequences. These protein motifs have evolved to recognize and fit into the grooves and contours of the DNA helix. Different DNA shapes, resulting from variations in base pair stacking, minor groove width, or DNA bending, can provide distinct structural features that proteins can recognize [32, 43].

During the catalytic cycle, enzymes frequently experience substantial conformational changes [44]. Several enzymes feature flexible loops or domains that can change conformation in response to substrate binding as a result of evolution. These conformational alterations may serve to isolate the active site from the surrounding area, resulting in a hydrophobic setting that is more catalytically favourable. In allosteric regulation, a molecule is able to switch between conformations because the binding of the regulatory molecule can either stabilize or destabilize the protein's active conformation. This conformational change is possible because proteins are adaptable molecules, that can take on several conformations [45–47].

1.4 Recognition and Repair of DNA Damage by Proteins

Disruptions to the genetic code or DNA can have a variety of consequences. Since proteins are the fundamental constituents of cells, changes or mutations in the DNA sequence can change their functionalities. If a gene that codes for a protein is mutated, the protein's structure and function can be altered, which can lead to various diseases. In some cases, mutations can be advantageous and give rise to new traits or adaptations. But mutations are mostly harmful and can result in a variety of genetic illnesses like cancer and Huntington's disease. This disruption or damage in DNA can be caused by environmental and natural agents that include UV radiation, ionizing radiation. DNA repair proteins are needed to correct these DNA damages, if left unattended the above mentioned repercussions can come true. Understanding the molecular specifics of the way these repair proteins identify and fix specific DNA lesions with higher precision is still fascinating, and is among the conundrums in this field of study because in order to maintain the integrity of the genome, efficient DNA repair proteins are required.

Despite the body's numerous fail safes and preventative measures, DNA can be damaged by factors such as UV light. UV light can damage DNA by generating pyrimidine dimers, which are covalent bonds between adjacent pyrimidine bases (thymine or cytosine) [48]. As these lesions disrupt the inter-base hydrogen bonding arrangements, these dimers have the ability to kink the DNA strand, which would obstruct normal DNA transcription and replication. The body has mechanisms in place to detect and repair such damage, proteins such as XPC (Xeroderma pigmentosum group C) and XPA (Xeroderma pigmentosum group A) are specifically involved in the recognition and repair of UV-induced DNA damage. XPC is a DNA damage recognition protein that binds specifically to damaged DNA.

Several mechanisms to repair the damaged DNA have been postulated, such as MMR (mismatch repair), which fixes abnormalities made during DNA replication like the insertion or deletion of nucleotides that can lead to base mismatches [49]. Base excision repair, or BER, is a process that repairs bases in DNA that have been damaged or altered. These modifications might be the result of chemical exposure or oxidative damage. Nucleotide excision repair, or NER, fixes substantial DNA damage brought on by environmental factors including UV radiation, which can generate DNA lesions like thymine dimers [50].

In order for DNA repair mechanisms to function properly, several proteins and enzymatic reactions must be precisely coordinated. Any interruption in these processes can result in insufficient or improper repair, resulting in genomic instability and potentially dangerous mutations. Additionally, many of the proteins involved in DNA repair are also involved in other cellular processes, and their malfunction can result in a wide range of human disorders. Thus, eukaryotic cells' ability to survive and function properly depends on sustaining adequate DNA repair. So, getting an insight into the repair mechanisms is an invaluabe asset.

Researchers have historically tried to create experimental ensembles of biomolecular conformations and they do provide valuable insights into the behavior of proteins and nucleic acids in vitro. Nevertheless, they can only offer a static image of the molecule at a specific time and they might not fully capture the variety of conformations that the molecule can assume since the flexibility and dynamics of the molecule are intrinsically challenging to observe in an experimental setting, especially during repair mechanism as the time scales for major events in DNA repair are very small. In this situation, computer modelling and simulation can be quite beneficial [51, 52].

The behaviour of biomolecules at the atomic or molecular level as a function of time can be captured in great detail using computational approaches like Monte Carlo and MD simulations [7]. They offer a thorough understanding of biomolecule behaviour that is challenging or impossible to get experimentally. And as every part of the biomolecule is simulated as a function of time, researchers can focus in on the molecular specifics that control how a system functions by concentrating on a particular region of interest, such as a protein active site or a DNA binding site.

To better understand the free energy gradients that control how biomolecules behave, several computational techniques have been created [9]. Free energy is a measure of the thermodynamic stability of a system and is related to the probability of observing a particular conformation or state. The structural and energetic parameters that control a biomolecule's behaviour can be understood by computing the free energy landscape of the molecule. The energy barrier in a free energy profile known as the activation barrier prevents a biomolecule from going through a certain process, such as a chemical reaction or a conformational change. And by calculating the free energy changes associated with the transition state of the process, researchers can gain insights into the energetic factors that govern the activation barrier. So, one can also learn more about the energetic parameters that control the activation barrier by computing the changes in free energy associated with the transition state of the activity.

From the antiquated methods like X-Ray crystallography, (late 1950s) to the more modern methods like NMR spectroscopy (1980s), experimental methods are utilised to ascertain the structure of proteins and nucleic acids [53–56]. They have made it possible to determine the complex molecular structure for more than 20,000 ligands, above 3000 unbound or complexed nucleic acids, and around 120,000 proteins that are grouped in the protein data bank (PDB) [57, 58]. Knowing the molecular structure of a biomolecule can provide insights into the biological processes in which it is involved, which can help in the development of new drugs and therapies. Accurate visualization of protein structures can aid in the design of drugs that specifically target and interact with the protein of interest. It can also help in identifying binding sites for other molecules, such as small molecule drugs or other proteins. But many proteins are not static, rigid structures and can have multiple conformations during a biological process that are necessary for their function. Furthermore, protein structures that have been crystallised or otherwise intentionally altered to aid in structure determination are the ones available in the PDB. So, the structures available at the PDB bank do not provide the complete picture and should be used with caution.

In order to accurately simulate a biological system, it is important to refine and correct the structures retrieved from databases, such as the Protein Data Bank (PDB), and to prepare the system for simulation. Correction of structural flaws, addition of missing residues, solvation of the system, addition of counterions, capping of terminal residues, and prediction of hydrogen atom locations are all part of this refinement process [59]. Force-field functions are used to determine the forces acting on each atom in the system following the modelling of the structures and during simulation. The interactions between the atoms in a molecule are represented mathematically by force fields. Bond stretching, bond angle bending, torsional rotations, non-bonded interactions like van der Waals forces, and electrostatic interactions are some of these interactions. And in these simulations, each atom in the system is treated as a point particle, and classical Newtonian mechanics is used to calculate the accelerations of each atom. The positions and velocities of the atoms are updated over time based on the forces acting on them, which are calculated using force-field functions mentioned above. Also, in these simulations, each atom in the system is modelled as a point particle, and the accelerations of individual atoms are determined using classical Newtonian mechanics [51]. The forces acting on the atoms are determined using the previously mentioned force-field functions, and these forces are used to update the atoms' locations and velocities over time.

In order to speed up calculations, high-performance computing clusters often include a large number of discrete processing units, such as CPUs and GPUs. These systems' parallelism, which is a vital component, enables researchers to carry out intricate simulations and calculations that would be impossible on a single machine [60–62]. Researchers generally utilise customised software packages that are tailored for these systems to take advantage of the parallelism built into high-performance computing clusters [63].

Biomolecular systems can have extremely irregular and complicated free energy landscapes, with numerous local minima corresponding to various conformations or states of the system. This complexity makes it difficult to obtain an accurate representation of the system's behavior through simulations, even with extended simulations [64]. This problem has been addressed by the development of a range of enhanced sampling methods like umbrella sampling [65] that can speed up conformational sampling and boost the effectiveness of molecular simulations. To encourage transitions between various conformations or states, these strategies often require altering the forces acting on the system's atoms.

1.5 Research Focus

UV-induced lesions are gaining a lot of attention because of their association with various types of cancer and other skin related diseases [48, 66]. Cyclobutane pyrimidine dimers (CPDs) and 6,4-

photoproducts (64PPs) are the 2 most prevalent UV-induced lesions. CPD is formed by covalent linkage of two adjacent pyrimidine bases (usually cytosine or thymine) in DNA. In CPD, C5 and C6 carbons of one pyrimidine base links to the C5 and C6 carbons of the adjacent pyrimidine base forming a cyclobutane ring structure that causes a distortion in the DNA double helix. On the other hand, 64PP is formed when adjacent pyrimidine bases in DNA become crosslinked together after exposure to UV radiation. In 64PP, a covalent bond forms between the C6 and C4 carbons of the two adjacent pyrimidine bases [67, 68]. Nucleotide excision repair (NER) is the primary method for repairing UV-induced DNA damage, and Xeroderma pigmentosum C (XPC) is the main protein responsible for discovering DNA damage [69–71].

In order to understand the pathways behind Xeroderma Pigmentosum (XP), a rare genetic condition characterised by excessive sensitivity to UV radiation and an elevated risk of skin cancer, it is crucial to investigate the molecular mechanism of XPC-mediated DNA damage repair [48, 72, 73]. Nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography are potent methods for figuring out the three-dimensional structures of macromolecules like DNA and proteins, as well as their complexes. In order to understand the precise interactions between the DNA and protein components as well as the conformational changes and dynamic processes involved in the repair process, researchers must first solve the structures of DNA-protein complexes that are involved in DNA repair. These structural studies can provide important initial clues into the molecular mechanisms of DNA repair. But the 3-D structure of the complex formed between damaged DNA and XPC protein has not been solved using the aforementioned techniques. The lack of a high-resolution structure for the XPC-DNA complex has limited our comprehension of the molecular mechanisms underlying its DNA repair activity [74–78]. A key component of the nucleotide excision repair (NER) process in yeast cells is played by the DNA repair protein Radiation sensitive 4 (Rad4). It is the XPC protein's yeast counterpart, which has a similar purpose in the NER pathway as Rad4 and XPC share significant sequence and structural similarity. Moreover, since DNA-Rad4 crystal structure is available, it can be utilised in place of XPC to study XPC-mediated DNA damage repair in humans [79–82].

During NER, the initial recognition of DNA damage is performed by a complex of proteins. Inducing a localised distortion of the DNA structure, this complex binds to the DNA lesion and aids in attracting more NER factors to the site of damage. Despite attempts at gaining experimental understanding, there is still much to learn about how these deformations are caused, and in turn, how the damaged DNA is recognised.

The crystal structure of RAD4-DNA complex consists of a DNA with UV-induced lesion (CPD or 64PP) is bound to Rad4 [83, 84]. Rad4 consists of an N-terminal transglutaminase domain (TGD), and 3 β-hairpin containing domains (BHD1, BHD2, and BHD3) [85, 86]. TGD and BHD1 attach to an 11-base pair section of duplex DNA that is undamaged, whereas BHD2 and BHD3 connect to a four base pair region or simply lesion site on the damaged strand of the DNA, making these domains essential for early damage detection [87, 88]. The flipped-out partner bases from the undamaged DNA strand are kept at the BHD2-BHD3 binding interface while the β-hairpin of BHD3 inserts into the DNA [84, 89–91].

It can be inferred from the crystal structure that during the early stage of lesion recognition by Rad4 by NER pathway, the important events are the association of RAD4 with DNA, flipping out of the lesion partner bases and BHD3 β-hairpin insertion into DNA duplex.

However, much remains to be discovered about the energetics and order of these events in a DNA containg CPD (Chapter 3) or 6-4PP (Chapter 4) with matched partner bases. This thesis presents an outlook on the mechanism and order of events during lesion recognition by XPC/Rad4 of DNA damage through an energetics perspective. The dynamics, structural changes and energetics of the Rad4-DNA complex are extensively studied in the study using molecular dynamics (MD) simulations and enhanced free energy sampling techniques like umbrella sampling [65].

Chapter 2

Computational Methods

2.1 Introduction

Computational techniques and algorithms are used in computational biology to understand biological systems and processes. This field of study takes concepts from biology, mathematics, computer science, and physics and applies them to a broad range of biological problems, from molecular to species-wide. Critical insights into the important processes occurring in dynamic, developing molecular systems can be obtained by using a structured analysis of molecular characteristics .

In the past it was thought that proteins have a single, rigid, 3-D structure and very little conformational flexibility. However, as technology and computational techniques started to advance, it became more and more clear that proteins were much more dynamic and flexible than previously thought. It was discovered by scientists that proteins could adopt numerous and distinct conformations which are necessary for protein function. With this understanding of flexibility new insights were gained into the relationship between structure and function as well as a thorough understanding of many biological processes. Due to these technological and computational advances we now have a much better understanding of the 3-dimensional spatial arrangement of amino acids in proteins. Now scientists are able to target the conformations of proteins involved in disease processes in order to create new drugs and therapies.

One such computational method is Molecular dynamics (MD) simulation, which is used to simulate the movements and interactions of molecules in a biological system. In MD, a model of the physics driving the interactions between the constituent particles is used to determine how each particle in a biomolecular system, which is ususally huge, moves with respect to time[7, 92]. Combining MD with experimental techniques such as such as X-ray crystallography and NMR spectroscopy provides a more comprehensive understanding of biological systems[93]. Scholars can better understand the molecular mechanisms behind biological processes, such as protein-protein interactions, ligand binding, and conformational changes in biological systems even on the scale of femtoseconds, by combining the advantages of experimental and computational methodologies.

We go into more detail about the basic premises and elements of molecular dynamics in this chapter, including the different algorithms, force fields, and analysis methods that are employed, with an emphasis on the elements that help us study protein-DNA and protein-ligand dynamics involved in critical cellular mechanisms.

2.2 Mathematical Formulation of Molecular Interactions

It is possible to predict the stability, reactivity, and kinetics of biomolecules and their connections using statistical mechanics, which is a valuable tool for examining the kinetics and thermodynamic properties like internal energy, heat capacity or pressure internal energy of biomolecular processes at the molecular level[94]. The momenta and locations of the N particles that make up the system of interest create the physiochemical characteristics of a many-body system. These properties can be described by the distribution of momenta and positions in the system, known as phase space. The system changes over time as the particles interact with one another, and the motion of individual particles is modelled using the laws of classical mechanics. In order to determine the time-averaged values of an interesting characteristic $\mathcal{A}$, it is required to design a realistic model of both the intra- and intermolecular interactions that occur within the system. Therefore, if at time t, $\{\mathbf{p}(t)\}$ represents N momenta and $\{\mathbf{r}(t)\}$ represents N positions, instantaneous property $\mathcal{A}$ can be defined as $\mathcal{A}(\{\mathbf{p}(t)\},\{\mathbf{r}(t)\})$. The positions (r) and momenta (p) of the individual particles in a biological process will change as they interact with one another which leads to change in instantaneous $\mathcal{A}$. By observing how the locations and momenta of the individual particles have changed, $\mathcal{A}$ averaged over time can be calculated as shown in following equation:

$$\langle \mathcal{A} \rangle_{\text{time}} = \lim_{\tau \to \infty} \int_{t=0}^{\tau} dt \, \mathcal{A}(\{\mathbf{p}\},\{\mathbf{r}\}) \tag{2.1}$$

If the integration is performed for τ approaching infinity, the value of the preceding integral approaches the ensemble value of $\mathcal{A}$.

2.2.1 Fundamentals of Statistical Mechanics

A system's thermodynamic state is typically characterised by its temperature T, pressure P, and volume V [94, 95]. These three parameters, along with the number of particles in the system (N) define the thermodynamic state of the system, and can be used to calculate its thermodynamic properties such as the internal energy, enthalpy, entropy, and free energy. The precise locations ($\{\mathbf{r}\}$) and momenta ($\{\mathbf{p}\}$) of each of the system's N constituent particles (N) determine the exact configuration of a system at the microscopic level. In statistical mechanics, 6N values are necessary to establish the microstate of a system with N atoms. This is because each atom has three position coordinates (r_x, r_y, r_z) and three momentum coordinates (p_x, p_y, p_z), so the total number of values required is 3N position values and 3N momentum values, or 6N total values. An ensemble can be thought of as a collection of points in phase space since each of its combinations represents a point in the 6N dimensional phase space. The following Hamiltonian equations govern phase space motion [96].

$$\begin{aligned} \frac{d\mathbf{r}_i(t)}{dt} &= \frac{\partial \mathcal{H}}{\partial \mathbf{p}_i} = \frac{\partial \mathcal{K}}{\partial \mathbf{p}_i} \\ \frac{d\mathbf{p}_i(t)}{dt} &= -\frac{\partial \mathcal{H}}{\partial \mathbf{r}_i} = -\frac{\partial U}{\partial \mathbf{r}_i} \end{aligned} \tag{2.2}$$

The range of i here being 1 to N. The total energy of a system for a given microstate is given by the Hamiltonian ($\mathcal{H}(\{\mathbf{p}\},\{\mathbf{r}\})$), which is the sum of all potential and kinetic energies of the particles in the system. In classical mechanics, the Hamiltonian ($\mathcal{H}(\{\mathbf{p}\},\{\mathbf{r}\})$) for a system of N particles can be written as the sum of potential ($U(\{\mathbf{r}\})$) and kinetic ($\mathcal{K}(\{\mathbf{p}\})$) energy. The total kinetic energy is calculated as a function of the momenta of the constituent particles by adding up their individual kinetic energies. The potential energy, on the other hand, is dependent on the locations of each pair of particles in the system and is calculated by adding the potential energies of each pair.

The concerned system is replaced by a great deal of replicas of the system in Gibbs' ensemble version of statistical mechanics, each with its own microstate, but all are maintained at the same thermodynamic conditions (such as temperature, pressure, and chemical potential). This allows the calculation of ensemble averages, which provide macroscopic information about the system, such as the average energy or the average number of particles in a given state. The ensemble average estimates the behaviour of a system as a whole rather than a single microstate and ensemble average of the property $\mathcal{A}$ can be calculated using the formula:

$$\langle \mathcal{A} \rangle_{ens} = \iint d^N\mathbf{p}\, d^N\mathbf{r}\; \mathcal{A}(\{\mathbf{p}\},\{\mathbf{r}\})\; \rho(\{\mathbf{p}\},\{\mathbf{r}\}) \tag{2.3}$$

Equation 2.3's double integral denotes 6N integral signs, for each of the system's atoms' 6N positions and momenta. Integrating over all potential system configurations yields the ensemble average of the property $\mathcal{A}$. $\langle \mathcal{A} \rangle_{ens}$, which will be termed to as the ensemble average henceforth, shows the average value of $\mathcal{A}$ across all ensemble replicas created during the simulation. The "expectation value" of a random variable is shown in angular brackets and the probability density of discovering an atom arrangement with momenta $\{\mathbf{p}\}$ and positions $\{\mathbf{r}\}$ is indicated by $\rho(\{\mathbf{p}\},\{\mathbf{r}\})$. To find the value of $\rho(\{\mathbf{p}\},\{\mathbf{r}\})$ we look towards Boltzmann distribution.

The probability density of a system with a constant particle number (N), volume (V), and temperature (T) is given by the Boltzmann distribution. The Boltzmann distribution is a statistical formula that describes the probability of a system being in a certain energy state, and is defined as:

$$\rho(p_i, r_i) = \frac{1}{Z} \exp\left\{ -\frac{1}{k_B T} E_i \right\} \tag{2.4}$$

In this context, $\rho(p_i, r_i)$ represents the probability of the system occupying energy state i. The partition function Z, Boltzmann constant k_B, energy of state E_i, and temperature T are incorporated into the equation. For the above mentioned canonical ensemble partition function, Z can be renamed as Q_{NVT} and since the temperature is constant $\frac{1}{k_B T}$ can be replaced with a contant β. The total energy corresponding to $\rho(\{\mathbf{p}\}, \{\mathbf{r}\})$ can be given by $\mathcal{H}(\{\mathbf{p}\}, \{\mathbf{r}\})$ which transforms the Equation 2.4 into following:

$$\rho(\{\mathbf{p}\}, \{\mathbf{r}\}) = \frac{1}{Q_{NVT}} \exp\{-\beta\mathcal{H}(\{\mathbf{p}\}, \{\mathbf{r}\})\} \tag{2.5}$$

The Hamiltonian of the system, denoted by $\mathcal{H}$, is often used to formulate the partition function, commonly abbreviated as Z. The partition function in the canonical ensemble, applicable to a system of N identical particles, can be written as follows:

$$Q_{NVT} = \frac{1}{N!}\frac{1}{h^{3N}} \iint d^N\mathbf{p}\, d^N\mathbf{r} \exp\{-\beta\mathcal{H}(\{\mathbf{p}\}, \{\mathbf{r}\})\} \tag{2.6}$$

In the Equation 2.6 $\frac{1}{h^{3N}}$ (here h is a dimensionality constant) is added to keep the partition function dimensionless whereas N! is present as N particles cannot be distinguished from one another and all possible configurations are taken into account. A different way to represent the partition function is to use the isobaric-isothermal ensemble's canonical partition function, which keeps the number of particles (N), pressure (P), and temperature (T) constant. Equation defining Q_{NPT} in terms of Q_{NVT} is as follows:

$$Q_{NPT} = \int dV\, Q_{NVT} \frac{\exp\{-\beta PV\}}{V_0} \tag{2.7}$$

2.2.2 Simplifying Free Energy Expression

The amount of energy a system has available to do work at constant temperature Free energy, often known as Gibbs free energy, is T. It is calculated by deducting the system's internal energy from the product of its temperature and entropy [97]. The transition in a system's free energy (ΔG) is related to the equilibrium constant of a chemical reaction and can be used to anticipate whether or not a reaction will be spontaneous and even the reaction rates can be calculated from the free energy barriers using the Transition State Theory. It is based on the idea that the rate-determining step in a reaction is the transition from the reactants to the products, through a highly energetic intermediate called the transition state[98]. The average time it takes for a system to cross the transition state and go from reactants to products (which is also equal to reciprocal of reate constant) is given by :

$$\tau = \frac{h}{k_b T} \exp\left(\frac{\Delta F}{k_b T}\right) \tag{2.8}$$

Gibbs free energy (ΔG), a measure of the system's potential for doing work at constant temperature and pressure with the condition of constant number of particles N is related to Q_{NPT} by the equation:

$$\langle G\rangle = \langle E\rangle - TS + PV = -\frac{1}{\beta}\ln Q_{NPT} \tag{2.9}$$

Here, V is the system's volume, S is its associated entropy, and $\langle E\rangle$ is its average internal energy.

Whereas Helmholtz free energy, A, which indicates the work that a system can perform under constant temperature and volume conditions, which may be used to compute the Gibbs free energy, is connected to the canonical partition function Q_{NVT} through the equation:

$$A = -\frac{1}{\beta}\ln Q_{NVT} \tag{2.10}$$

As mentioned earlier $\mathcal{H}$ is the total energy of system and total energy constitutes of two energies kinetic and potential. So, by replacing the $\mathcal{H}$ in Equation 2.6 by kinetic and potential energies we get:

$$\begin{aligned} Q_{NVT} &= \frac{1}{N!}\frac{1}{h^{3N}}\iint d^N\mathbf{p}\, d^N\mathbf{r}\exp\{-\beta\mathcal{H}(\{\mathbf{p}\},\{\mathbf{r}\})\} \\ &= \frac{1}{N!}\frac{1}{h^{3N}}\iint d^N\mathbf{p}\, d^N\mathbf{r}\exp\{-\beta[\mathcal{K}(\{\mathbf{p}\}) + U(\{\mathbf{r}\})]\} \\ &= \frac{1}{N!}\frac{1}{h^{3N}}\int d^N\mathbf{p}\exp\{-\beta\mathcal{K}(\{\mathbf{p}\})\}\int d^N\mathbf{r}\exp\{-\beta U(\{\mathbf{r}\})\} \\ &= \frac{1}{N!}\frac{1}{h^{3N}}\prod_{i=1}^{N}\int d\mathbf{p}_i\exp\left\{-\beta\frac{\|\mathbf{p}_i\|^2}{2m_i}\right\}\int d^N\mathbf{r}\exp\{-\beta U(\{\mathbf{r}\})\} \end{aligned} \tag{2.11}$$

After isolating the product of integral in the above equation we get:

$$K = \prod_{i=1}^{N}\int d\mathbf{p}_i\exp\left\{-\beta\frac{\|\mathbf{p}_i\|^2}{2m_i}\right\} \tag{2.12}$$

Here, each component of the product is separately evaluated to the following:

$$K_i = \left(\sqrt{\frac{m_i}{2\pi\beta}}\right)^3 \tag{2.13}$$

Now the integrated value of the product of the integrals can be substituted in Equation 2.11 and after normalising all constants in it we finally get the simplified version of the equation as mentioned below:

$$Q_{NVT} = \int d^N\mathbf{r}\ \exp\{-\beta U(\{\mathbf{r}\})\} \tag{2.14}$$

Similarly, Q_{NPT} can be simplified as well and we get the following equation:

$$Q_{NPT} = \frac{1}{C}\int dV\ \exp\{-\beta E(N,V)\} \tag{2.15}$$

Here, E is the internal energy of the system, V is the volume, N is the number of particles and C accounts for the normalization of the integral.

2.2.3 Potential Mean Force (PMF)

The effective interaction potential between two or even more particles in a system is described by the theoretical concept known as the potential of mean force or PMF. It is a measurement of the average force a particle experiences while in a specific arrangement, accounting for the influence of all other particles in the system. The PMF is a way of quantifying the thermodynamic driving force for a system to move from one state to another. A multidimensional reaction coordinate (η) is typically defined to conduct studies on free energy differences and calculate PMF. By varying the multidimensional reaction coordinate, the system's energy can be altered and the PMF can be calculated for different thermodynamic states. It is calculated as the negative derivative of the logarithm of the marginal probability distribution of a single particle's coordinates, such as its position or its volume:

$$\mathrm{PMF}(\eta) = -\ln(Q(\eta)) \tag{2.16}$$

Where η is the coordinate of interest, and $Q(\eta)$ is the marginal probability distribution of η. This distribution can be calculated from the full configuration of the system by integrating over all other degrees of freedom, such as the positions ($\{\mathbf{r}\}$) and momenta ($\{\mathbf{p}\}$) of the other particles.

As shown in Equation 2.5 gives the probability distribution for the system to exist in a specific sequence of $\{\mathbf{r}\}$ and $\{\mathbf{p}\}$. As mentioned before by integrating over all other degrees of freedom we get the final probability distribution:

$$Q(\eta) = \frac{\int \delta[\eta(\{\mathbf{r}\}) - \eta] \exp\{-\beta U(\{\mathbf{r}\})\} d^N\mathbf{r}}{\int \exp\{-\beta U(\{\mathbf{r}\})\} d^N\mathbf{r}} \tag{2.17}$$

2.2.4 The Ergodic Hypothesis

A fundamental idea in statistical mechanics known as the ergodic hypothesis states that, if given enough time, a system will ultimately travel to all of its accessible microstates. However, it's important to note that MD simulations have limited time and length scales, so it is not always possible to sample the entire phase space. This means that the trajectory generated by the simulation may not be truly ergodic, or representative of the behavior of the system over an infinite time even if MD simulations are designed to sample the phase space of a system, which is the space of all possible positions and momenta of the system's particles, in order to approximate the thermodynamic properties of the system [94, 99].

The system is modelled as a large number of particles that interact with each other over time in Mollecular Dynamics simulations. Calculating different quantities of concern, such as temperature, pressure, or potential energy, can be used to study the system's behaviour. A quantity's time average is calculated by observing the system's behaviour over a specified period and averaging the quantity's values over that time frame. The ensemble average of a quantity, in contrast, is calculated by running multiple independent simulations of the system, each with a different set of initial conditions, and averaging the quantities' values across the ensemble of simulations. Because the findings of an MD simulation rely heavily on the length of the simulation and the initial values chosen, it is vital to establish a relationship between both the time average and the ensemble average of items of interest. It is possible to see how MD simulation results depend on the length of the simulation and the initial conditions by relating the time average to the ensemble average. This is crucial for comprehending the simulation results' accuracy and limitations. So, finding a relation between time average and ensemble average calculations of quantities of interest is important in Mollecular Dynamics simulations because it helps to understand the accuracy and limitations of the simulation results. But as mentioned before it is difficult for a MD simulation to traverse whole phase state and provide a true ensemble average.

As a workaround to calculate the true ensemble average, an estimation to the ensemble average of a particular physical quantity which is a function of momenta and positions over the microcanonical ensemble is discovered by calculating that quantity at each timestep and averaging.

$$\langle \mathcal{A} \rangle_{\tau \to \infty} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} \mathrm{d}t \, \mathcal{A}(\{\mathbf{p}(t)\}, \{\mathbf{r}(t)\}) \tag{2.18}$$

This is possible because the system transitions from one microscopic configuration to another at evey timestep in Molecular dynamics. This microscopic configuration is unmistakably a representation of a microstate in the microcanonical (constant N, V, and E) ensemble because the equations of motion (Hamilton's equations [96]) and their arithmetical application in MD (e.g., the velocity Verlet algorithm [100]) are energy conserving with number of particles and volume remaining constant and as a result calculating, averaging a physical quantity which is a function of momenta and position works in estimating ensemble average. In light of the fact that the time average in an MD simulation can indeed be taken as the ensemble average, one can even calculate free energy using MD. The integral mentioned in Equation 2.18 is interpreted by summing up the product of integrand and integrator as Δt because simulations are performed in discrete time, and since computation time is relatively limited, the integral's upper bound is fixed.

2.3 Force Fields

The Born-Oppenheimer approximation and force field methods are used to calculate the potential energy of a molecular system. In this approximation, energy is viewed as a function of atomic nuclear coordinates while electronic mobility is ignored [94]. TA set of empirical or semi-empirical potentials is used to explain the interactions between atoms, including bonds, angles, torsions, and van der Waals interactions. The structure and energetics of a molecule, as well as how it will react to different environmental perturbations like temperature, pressure, and solvation, can all be predicted using force field methods by calculating the potential energy of a molecular system as a function of its atomic nuclear coordinates.

The Born-Oppenheimer approximation plays a crucial role in making force field methods calculations simpler, enabling calculations on large systems with large quantities of atoms. When the motion of a molecular system's nuclei and electrons are separated, it is much simpler to calculate the potential energy of the system. This makes it possible to use efficient computational methods to study large and complex molecular systems, as opposed to quantum mechanics, where one must take into account all of the electrons, greatly increasing the complexity [101]. Modern force fields can perform calculations that are as accurate as some of the most complex quantum mechanical calculations while vastly reducing the computational cost. The force field computation accuracy has significantly increased over time [102]. It's important to remember that not all force fields are equally accurate for all types of molecular systems, and that the choice of force field still influences how accurate force field calculations are. The empirical or semi-empirical nature of force field methods continues to be a drawback, they may not always accurately capture a molecular system's behavior under challenging circumstances, such as when covalent bonds are broken or for highly reactive species.

A model of "balls connected by springs" is used in the force field method. Spheres (or "balls") are used to represent atoms in this model, and springs are used to represent the interactions between atoms. The bond strengths and bond lengths between atoms are described, respectively, by the strength of the springs and their equilibrium lengths [103, 104]. Each interaction between the atoms is represented by a term in the potential energy function. The total of the atoms' bond, angle, dihedral, and non-bonded interactions is used to calculate the system's potential energy. In order to match experimental data, such as bond lengths, bond angles, and vibrational frequencies, or to match the outcomes of quantum mechanical calculations, the parameters of the potential energy function are altered. For many applications, the balls-connected-by-springs model is a good approximation of the behaviour of a molecular system because it is straightforward and intuitive in how it depicts the interactions between atoms in a molecular system. Since the model is straightforward and reasonably easy to implement in computer simulations, it is used in molecular dynamics simulations and other computational studies of molecular systems [105].

A 4-component characterization of the biomolecular forces within the system can be used to explain contemporary molecular modelling force fields, such as FF14SB [94, 107]. Those four componenets being:

- **Bond Stretching:** This component represents the stretching or compression of covalent bonds between atoms. The bond stretching component of the force field is typically described by a harmonic potential energy function, which accounts for the bond length and the bond strength. According to Hooke's Law, each bond stretch between two atom contributes the following amount to the overall molecular potential energy, assuming that the system is modelled after masses at-

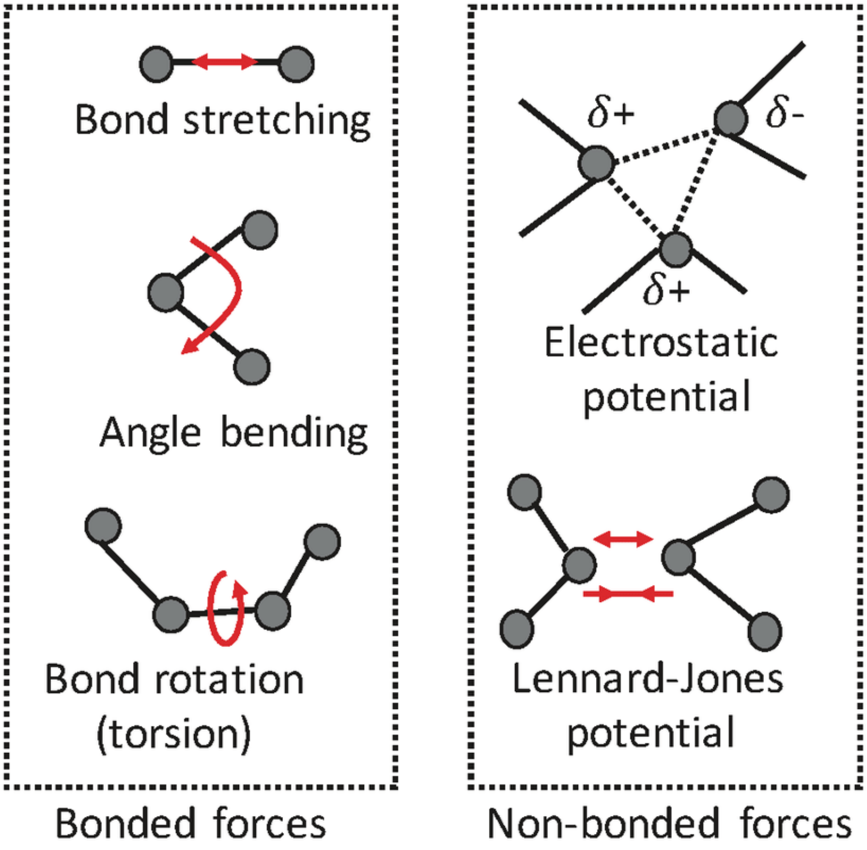


Figure 2.1: Here is a schematic representation of the various contributions to a force field energy calculation. As shown in the figure Bonded forces represent the first three componenets, bond streching, angle bond bending, and bond rotation in the Equation 2.23. The non-bonded forces represented here constitutes the fourth term in the aforementioned equation. Illustration and caption taken from [106]

tached to springs:

$$\sum_{\text{stretch}} U_{\text{bond},i} = \sum_{\text{bonds}} \frac{1}{2} k_{\text{bond},i}(r_i - r_{\text{eq},i}) \tag{2.19}$$

In the given equation, $U_{\text{bond},i}$ represents the energy associated with bond stretching, $k_{\text{bond},i}$ denotes the force constant of the bond, r_i represents the current bond length, and $r_{\text{eq},i}$ represents the equilibrium bond length for the i^{th} bond.

- **Bond Angle Bending:** This component represents the bending of bonds between atoms. The force field's bond angle bending element is typically described by a harmonic potential energy function:

$$\sum_{\text{bend}} U_{\text{angle},i} = \sum \frac{1}{2} k_{\text{angle},i}(\theta_i - \theta_{\text{eq},i}) \tag{2.20}$$

where $U_{angle,i}$ represents the energy associated with bond angle bending, $k_{angle,i}$ denotes the force constant of the bond angle, Θ_i represents the current bond angle, and $\Theta_{eq,i}$ represents the equilibrium bond angle of the i^{th} bond angle in the system.

- **Torsion (Dihedral) Angle:** This element symbolises the atom-to-atom bonds twisting. A multi-term potential energy function, such as dihedral motions around four bonded connected atoms and other out-of-plane inversion motions, is typically used to describe the torsion angle component of a force field.

$$\begin{aligned}\sum_{\text{torsion}} U_{torsion,i} = \sum_{\text{dihedral}} & k_{dihedral,i}(1 - \cos(n\omega - \gamma_{dihedral})) + \\ \sum_{\text{inversion}} & k_{inversion,i}(1 - \cos(n\psi - \gamma_{inversion}))\end{aligned} \tag{2.21}$$

  Where, $k_{dihedral,i}$, $k_{inversion,i}$, $\gamma_{dihedral}$, $\gamma_{inversion}$ are terms that need to be determined through experiments, and and ω and ψ are the dihedral and out-of-plane inversion angles, respectively.

- **Non-Bonded Interactions:** This component represents the interactions between atoms that are not bonded to one another, such as van der Waals interactions, electrostatic interactions, and hydrogen bonding. The Lennard-Jones potential [108] and the Coulomb potential are two instances of the empirical potential energy functions that are generally used to describe the force field's non-bonded interactions portion.

$$\sum_{\text{non}-\text{bonded}} U_{i,j} = \sum_{\text{coulomb}} \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{r_{i,j}} + \sum_{\text{vdW}} 4\varepsilon_{i,j} \left[\left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{12} - \left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{6} \right] \tag{2.22}$$

  According to the regulations of the specific force field mentioned here, atomic charges q_i and q_j are assigned to the i^{th} and j^{th} atoms, respectively. Since electrostatic interactions cannot fully explain all of the nonbonded interactions in a system, the van der Waals interactions are modelled using the Lennard-Jones potential [108]. A rare gas atom is a clear illustration of a system where dipole-dipole or dipole-induced dipole connections are not possible. But the presence of solid and liquid phases in rare gases as well as behavior that differs from that of an ideal gas suggests the existence of extra contacts between the atoms. Modeling of the interaction energy is shown in the Figure 2.2. The Lennard-Jones potential results from the equilibrium of attractive and repellent forces.

The functional form of the majority of common molecular force fields is obtained by adding up the aforementioned elements [94]:

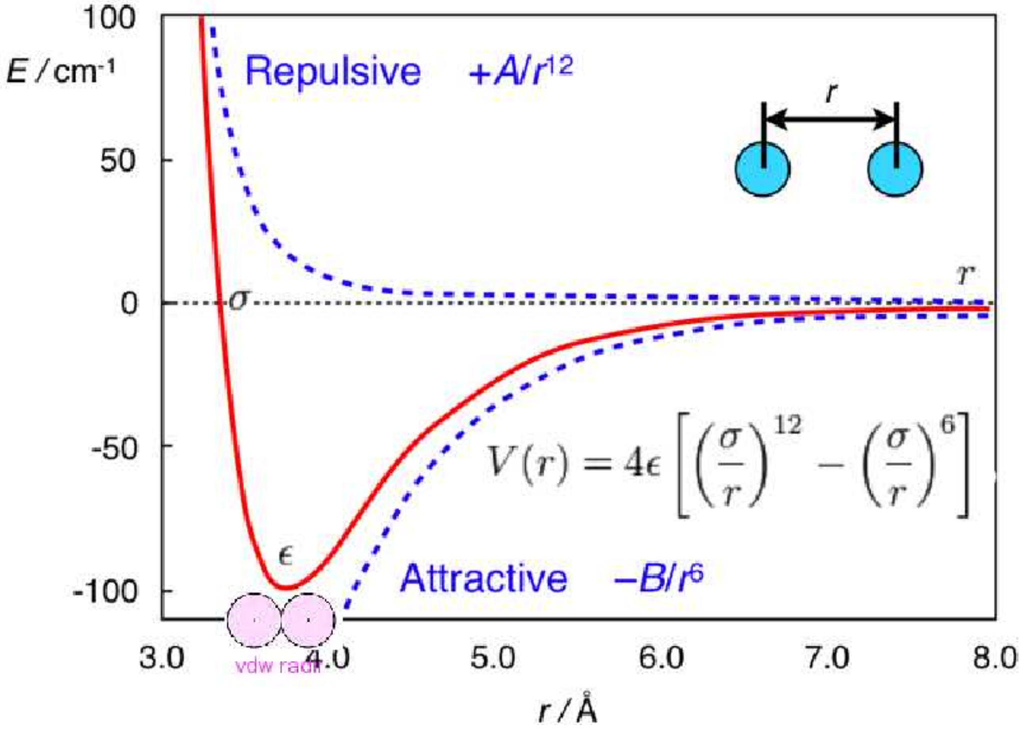


Figure 2.2: Functional form of the Lennard-Jones Potential $U(r)$ as a function of the interatomic distance r. Illustration and caption taken from [109]

$$U(\mathbf{r}^N) = \sum_{\text{stretch}} U_{bond,i} + \sum_{\text{bend}} U_{angle,i} + \sum_{\text{torsions}} U_{torsion,i} + \sum_{\text{vdW}} U_{i,j} + \sum_{\text{coulomb}} U_{i,j} \tag{2.23}$$

Where $U(\mathbf{r}^N)$ stands for potential energy as a function of a N particle system's positions ($\mathbf{r}$).

In conclusion, the four components of the biomolecular forces - bond stretching, bond angle bending, torsion angle, and non-bonded interactions can be used to describe the modern molecular modelling force fields. Together, these components make up a potential energy function that is used to calculate the forces and energies acting on the atoms as well as to describe the interactions between atoms in molecular systems.

2.4 Energy Minimization

2.4.1 Potential Energy Surfaces

The intermolecular potential energy function, $U(\{\mathbf{r}\})$ in Equation 2.2 which describes the interactions between the component particles in a molecular system, defines the potential energy surface in simulations of molecular dynamics as a function of system arrangement $(\{\mathbf{r}\})$. An essential presumption that is frequently applied in molecular dynamics (MD) simulations is the Born-Oppenheimer approximation. It states that the motion of the nuclei in a molecular system can be treated as a slow motion relative to the motion of the electrons. As a result, it is possible to treat the motion of the electrons as a distinct, quicker motion and to treat the nuclei as being in a fixed configuration. Also, the nuclei positions can be used to model the electronic structure of a molecule [94]. A system's total energy comprises of two components: electronic energy and nuclear energy. Nuclear energy is a function of nuclei positions and momenta, whereas electronic energy is a function of nuclei positions and electron wave functions. So, a molecule's energy in its electronic ground state is simply a function of its nuclear coordinates. The flow of a large group of atoms as the molecule transitions from one conformation to another or even simple and direct procedures like translation, rotation and vibration can cause changes in the positions of atomic nuclei. The system's energy typically changes depending on the type of change. For two different systems even simple changes like increasing the bond length of cvovalent bonds can require completely different amounts of energy [94].

A large systems' potential energy is commonly described as a complex multidimensional function, implying that it depends on a wide range of factors. This is because any given system's potential energy is determined by interactions between its atoms or molecules, which can be incredibly complicated and multidimensional in nature. Take a large protein molecule as an illustration, which may have thousands of atoms. The potential energy of this system depends on the positions and orientations of each atom in the protein as well as on how they interact with the surrounding solvent, such as water. The electrostatic

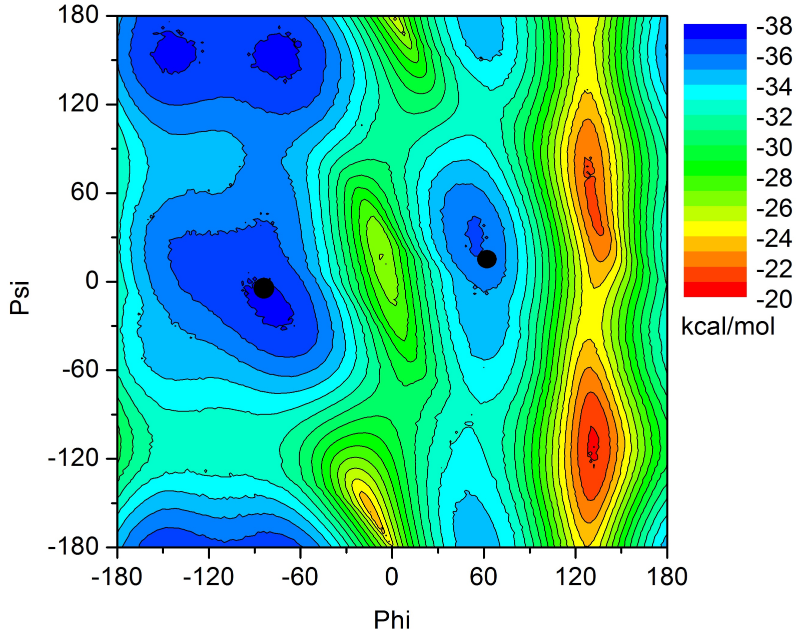


Figure 2.3: The potential energy surface is represented as a 2D contour as a function of the Ψ and Φ, the backbone dihedral angles, which identifies the rotation of a di-alanine peptide molecule about the $C_5 - C_6$ bond. Illustration and caption taken from `http://ambermd.org/tutorials/advanced/tutorial5_amber11/section2.htm`

interactions between charged atoms, the van der Waals interactions between non-bonded atoms, and the creation of hydrogen bonds between atoms are all components of the potential energy function used in a molecular dynamics simulation. In a system containing only di-alanine peptide molecule it is possible to study the energy variations during rotation of molecule along the $C_5 - C_6$ bond as a function of Φ and Ψ. A 2D contour that represents the energy values for all possible combinations of Φ and Ψ can be created by ranging the torsion angles from $0°$ to $360°$ and calculating the energy for each new conformation.

A system's position on the multidimensional energy surface changes when its energy changes, as does the system itself. The system moving from one place to another while moving along the energy surface can be used to represent this change in position. The energy of the system is constant with respect to the atomic coordinates at stationary points on the energy surface. On the energy surface, minimum points and saddle points are the two main categories of stationary points. The lowest points on the energy surface are the minimum points, which are stable configurations of the system. When the system reaches its minimum point, its energy is at its lowest point and will increase with even the slightest disturbance. This suggests that any variations to the system will need an energy input and that the system will typically remain in its lowest energy configuration. On the other hand, saddle points correspond to locations where local energy is greatest along one or more dimensions but is lowest along other dimensions. As the system is not stable at a saddle point, any slight disturbance will cause it to move away from it and towards one of the nearby minimum points. Saddle points are important in chemical reactions because they represent transition states, the highest energy configurations along the reaction pathway. The term "activation energy" refers to the system's energy at a saddle point and denotes the bare minimum of energy required to initiate the reaction.

2.4.2 Minimisation of energy surfaces

Finding the minimum value of a function f involves locating the point on the function that corresponds to the lowest value. At a minimum point, the function's first derivative equals zero. In other words, the function's gradient—a vector that points in the direction of the steepest ascent—is orthogonal to the surface at its minimum point. The second derivative of the function is positive at a minimum point, indicating that the function is locally concave.

In Molecular Dynamics the potential energy $U(\{\mathbf{r}\})$ of a molecular system is the function of great interest. By taking into account the sum of the interaction energies of all the atoms in the system, this scalar value gauges the stability and dynamics of the system over time. The positions and velocities of the atoms inside the system are the most crucial variables in MD simulations. These variables establish the system's configuration and are used to calculate the potential energy and other aspects of the system. The positions of individual atoms are described using a coordinate system called cartesian coordinates. In classical molecular mechanics, where the energy is determined by a function of $3N$ variables, this coordinate system is the one that is most frequently used for minimization.

Since most minimization algorithms can only go downslope, they can only find local minima rather than global minima, which is their main drawback. A local minimum is a location on the potential energy surface where the energy has a lower value than the neighbouring points. There may be another minimum with lower energy located elsewhere in the system. The minimum point in the system with the lowest energy is known as the global minimum. Even if there is a lower energy global minimum that has not yet been discovered, a minimization algorithm will halt when it reaches a local minimum. In order to ensure that the global minimum is found, it may be necessary to run the minimization several times from different starting points. This implies that the starting conditions may have an impact on the minimization process's outcomes. In order to overcome this disadvantage there are some special minimisation methods that can make uphill moves to seek out global minima rather than the nearest local minima. But here we'll examine two typical methods for achieving energy minimization: (a) Gradient Descent method and (b) the Conjugate Gradient method. Both the gradient descent method and the conjugate gradient method change the coordinates of the atoms in order to minimize the energy of the system. The user-provided starting arrangement of the system, known as $\{\mathbf{r}\}_0$, serves as the start for the inaugural iteration. Following that, the configuration obtained from the previous step, represented by $\{\mathbf{r}\}_{k-1}$, serves as the starting point at step k.

2.4.3 Gradient Descent

The coordinates are revised when employing the gradient descent method in the direction of the negative gradient, which is also the direction of the steepest descent. The technique moves slowly along the negative gradient, reassessing the system's energy at each step. Until the energy of the system stops fluctuating or reaches a minimum, this process is repeated. The algorithm begins with an initial parameter ($\{\mathbf{r}\}_0$) given by user and determines the cost function's ($U(\{\mathbf{r}\})$ in this case) gradient with respect to the parameters at that time. Moving in the opposite direction will cause the cost function to be minimised because the gradient indicates the direction of the steepest ascent. To accomplish this, a portion of the gradient - whose fraction depends on the step size is subtracted from the parameters. To ensure that the algorithm converges to the minimum, the step size, which determines the size of the update to the parameters, must be carefully chosen. As an equation it can be represented as:

$$\{\mathbf{r}\}_{n+1} = \{\mathbf{r}\}_n - \alpha_n \boldsymbol{\nabla} U(\{\mathbf{r}\}_n) \tag{2.24}$$

Where the step size is α_n and gradient of $U(\{\mathbf{r}\}_n)$ is $\boldsymbol{\nabla} U(\{\mathbf{r}\}_n)$.

2.4.4 Conjugate Gradient

The conjugate gradient method [110] is a function minimization algorithm. It is based on the gradient descent algorithm, which updates the values of the variables iteratively in order to decrease the function's value. The conjugate gradient method allows for more efficient convergence by selecting an update direction that is conjugate to the direction of the previous update. The mathematical formulation of the conjugate gradient method involves the following steps:

- Given a starting point, $\{\mathbf{r}\}_0$, and a function, $U(\{\mathbf{r}\})$, with gradient, $\boldsymbol{\nabla} U(\{\mathbf{r}\})$.
- Initialize the search direction, p_0, to be the negative gradient at the starting point, $p_0 = -\boldsymbol{\nabla} U(\{\mathbf{r}\}_0)$.
- Choose a step size, α, and update the values of x and p according to the following equations:

$$\{\mathbf{r}\}_{n+1} = \{\mathbf{r}\}_n + \alpha p_n \tag{2.25}$$

$$p_{n+1} = -\boldsymbol{\nabla} U(\{\mathbf{r}\}_{n+1}) + \beta p_n \tag{2.26}$$

where β is a scalar value that determines the direction of the search, and is computed using the following formula:

$$\beta = \frac{\nabla U(\{\mathbf{r}\}_{n+1})^{T \nabla U(\{\mathbf{r}\}_{n+1})}}{\nabla U(\{\mathbf{r}\}_{n})^{T \nabla U(\{\mathbf{r}\}_{n})}} \tag{2.27}$$

- Previous two steps should be repeated until a stopping criterion is reached, such as a specified maximum number of iterations or a tolerance for change in function value.

The conjugate gradient method is iterative, and with each iteration, the values of $\{\mathbf{r}\}$ and p are updated based on the gradient that is currently in effect and the previous search direction. Large optimization problems are well suited for the this method because it avoids retracing steps that have already been taken.

2.5 Molecular Dynamics

Molecule dynamics (MD), a method of computer simulation, is used to study the dynamics of molecules, including their motions and interactions [111]. The technique calculates the trajectory of the molecules as a result of the interactions between atoms and molecules using classical mechanics concepts, such as the laws of motion. Through microscopic simulations, it can be used to investigate the macroscopic properties of a system, for instance, to investigate the energetics and mechanisms of conformational change as shown in Chapter 3 and 4. By modelling a system of particles in motion, MD simulations provide a way to study the dynamic behavior of the system and the relationships between its thermodynamic and kinetic properties. [112].

Newton's equations of motion, which explain how a system of particles moves, are solved numerically in MD simulations. These equations, $F = ma$, relate the force acting on a particle to its mass (m) and acceleration (a). The gradient of the system's potential energy, which describes the interactions between the particles, governs the force acting on a particle in MD simulations.

Now imagine a system of N particles, each of which has the following properties:

- N particles
- Atomic coordinates: $(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) = \{\mathbf{r}\}$
-
- Atomic momenta: $(\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_N) = \{\mathbf{p}\}$
- Potential energy: $U(\{\mathbf{r}\})$
- Kinetic energy: $\mathcal{K}(\{\mathbf{p}\}) = \sum_{i=1}^{N} \frac{\|\mathbf{p}_i\|^2}{2m_i}$

Classical eauations of motions can also be written as:

$$\dot{\mathbf{r}}_i = \frac{\mathbf{p}_i}{m_i} \qquad \text{and} \qquad \dot{\mathbf{p}}_i = \mathbf{F}_i(t) \tag{2.28}$$

When written in this form, it is clear that the classical equations of motion are fundamentally a system of associated ordinary differential equations, where i ranges between 1 and N. Depending on the positions of all other particles in the system at that time, $\mathbf{F}_i(t)$'s force on each atom i at time t can be calculated as follows:

$$\mathbf{F}_i(t) = -\boldsymbol{\nabla}_i U(\{\mathbf{r}(t)\}) = -\boldsymbol{\nabla}_i \sum_{i=1}^{N} \sum_{j>i}^{N} U(r_{ij}(t)) \tag{2.29}$$

In this case, the gradient is calculated with respect to the position of atom i and $U(r_{ij}(t))$ is the selected pairwise potential between atoms i and j. Also, $r_{ij}(t)$ is equivalent to $\|\mathbf{r}_i(t) - \mathbf{r}_j(t)\|$. The fundamental MD algorithm repeats the following steps after the initial conditions are set:

1. The forces acting on each atom, as depicted in the Equation (2.29), are calculated using the potential energy function..

2. the atoms' velocities are modified based on the forces and the time step using the second part of Equation (2.28).

3. Based on the velocities and the time step, the atom locations are modified using Equation (2.28).

4. The simulation continues for a specified number of time steps or until a specific criteria is met (e.g., a certain temperature or pressure).

5. The system's numerous thermodynamic parameters, such as temperature, pressure, energy, and other pertinent values, can be calculated after each time step.

Repeating this cycle results in a time series of molecular positions and velocities, which can be used to determine the system's properties over time. To calculate velocities and positions mentioned in Step 2 and 3 integration of Equation (2.28) is required. But as the Molecular dynamics calculates them at small discrete time steps (ranging from 1 to 10 femtoseconds), The earliest "Verlet" algorithm [113, 114], which can be derived by taking into account the Taylor expansion concerning time step t of the location of the atom, can be used to get around the integral. The following are the equations:

$$\mathbf{r}_i(t+\Delta t) \;=\; \mathbf{r}_i(t) + \frac{\mathbf{p}_i(t)}{m_i}\Delta t + \frac{\mathbf{F}_i(t)}{2m_i}\Delta t^2 + \frac{1}{3!}\frac{d^3\mathbf{r}_i}{dt^3}\Delta t^3 + O(\Delta t^4) \tag{2.30}$$

By replacing $+\Delta t$ with $-\Delta t$ we get,

$$\mathbf{r}_i(t-\Delta t) \;=\; \mathbf{r}_i(t) - \frac{\mathbf{p}_i(t)}{m_i}\Delta t + \frac{\mathbf{F}_i(t)}{2m_i}\Delta t^2 - \frac{1}{3!}\frac{d^3\mathbf{r}_i}{dt^3}\Delta t^3 + O(\Delta t^4) \tag{2.31}$$

After adding both the above mentioned equations and subtracting $\mathbf{r}_i(t-\Delta t)$ on both sides, we get:

$$\mathbf{r}_i(t+\Delta t) \;=\; 2\mathbf{r}_i(t) - \; \mathbf{r}_i(t-\Delta t) + \; \frac{F_i(t)}{m_i}\Delta t^2 + \; 2O(\Delta t^4) \tag{2.32}$$

in which $O(\Delta t^4)$ is a 4-order error term.

2.5.1 The Velocity Verlet Integrator

Equation (2.32) does not explicitly calculate the momenta, but it can be used to update the locations of the N particles at every time step t. By connecting the Taylor expansion from the original Verlet algorithm, as shown in Equation (2.30) to the next technique for updating the momenta, the velocity Verlet algorithm [100] corrects this the following equation:

$$\mathbf{p}_i(t+\Delta t) \;=\; \mathbf{p}_i(t) + \; \frac{F_i(t+\Delta t) + F_i(t)}{2}\Delta t \tag{2.33}$$

The momenta of the particles are computed at the half-time step ($\frac{1}{2}\Delta t$) of the Velocity Verlet algorithm [100]. The momenta of the particles at a half time step are calculated using the initial positions and velocities of the particles. This is accomplished by using the particle's current velocity and updating it with the forces ($F_i(t)$) that are acting on it. The algorithm updates the particle positions for the upcoming full time step ($r_i(t + \Delta t)$) using the calculated momenta ($p_i(t + \frac{1}{2}\Delta t)$). From the updated positions, the forces acting on each particle are calculated ($F_i(t + \Delta t)$), and the updated forces and half-step momenta are used to update the momenta at the full time step. This can be summarised by the following equations:

$$\mathbf{p}_i(t + \tfrac{1}{2}\Delta t) = \mathbf{p}_i(t) + \tfrac{1}{2}\Delta t\, F_i(t) \tag{2.34}$$

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \Delta t\, \mathbf{p}_i(t + \tfrac{1}{2}\Delta t) \tag{2.35}$$

$$\mathbf{p}_i(t + \Delta t) = \mathbf{p}_i(t + \tfrac{1}{2}\Delta t) + \tfrac{1}{2}\Delta t\, F_i(t + \Delta t) \tag{2.36}$$

The Velocity Verlet method's ability to calculate each particle's position and velocity simultaneously at the same time step is one of its main features [100]. This is important because a particle's velocity depends on both its position and velocity at a given time step. The Velocity Verlet method [100] prevents any errors that might happen if only one quantity was updated at a time by updating both quantities simultaneously. All of the simulations conducted for this study have used the velocity Verlet algorithm due to this characteristic.

The force acting on each particle is typically calculated in molecular dynamics simulations using an inter-particle potential, such as the Lennard-Jones potential or the Coulomb potential. Evaluation of these potentials can be computationally expensive, particularly in large systems with numerous particles. The Velocity Verlet algorithm seeks to minimise the number of times the force is calculated in order to lessen the computational burden of the simulation, similar to other numerical methods for integrating the classical Newtonian laws. As mentioned before, the particle positions and velocities are updated at each time step using the forces and momenta computed at the half time step. The Velocity Verlet algorithm [100] can improve simulation efficiency by minimising the number of times the forces are calculated, enabling simulation of larger and more complex systems in a reasonable amount of time [59].

2.5.2 Extended Hamiltonian Methods

As shown above in traditional simulations of molecular dynamics, the system evolves over time in accordance with the Hamiltonian, which characterises the overall energy of the system corresponding to NVE ensemble. Nevertheless, the majority of biochemical functions take place in environments with constant volume (V), temperature (T), pressure (P), or maybe a pairing of these situations. By adding more terms to this fundamental Hamiltonian, the extended Hamiltonian approach extends it and enables the simulation to produce the desired ensembles like NPT or NVT.

For instance, in Berendsen thermostat implementation [115] in AMBER, an extended Hamiltonian contains a term that couples the system to a heat bath in order to produce the canonical ensemble (NVT), effectively regulating the system's temperature T_0. As a result, even when the system's energy changes as a result of particle interactions, the simulation can keep the temperature constant. This property led to its use in the majority of the simulations conducted for this study. Similar to this, the extended Hamiltonian contains terms that regulate the system's temperature and pressure in order to produce the isobaric-isothermal ensemble (NPT). The system's volume changes as a result of interactions between the particles, but the simulation is still able to keep the temperature and pressure constant.

2.6 Enhanced Sampling Techniques

Following the integration of all degrees of freedom $\xi(\{\mathbf{r}\})$, but ξ, we obtain the probability distribution of the system along ξ, as shown in Equation (2.17) and below:

$$Q(\xi) = \frac{\int \delta[\xi(\{\mathbf{r}\}) - \xi] \exp\{-\beta U(\{\mathbf{r}\})\} d^N\mathbf{r}}{\int \exp\{-\beta U(\{\mathbf{r}\})\} d^N\mathbf{r}} \tag{2.37}$$

In a canonical ensemble, the free energy $A(\xi)$ along the reaction coordinate ξ can be represented as shown in Equation (2.10)

$$A(\xi) = -\frac{1}{\beta} \ln Q(\xi) \tag{2.38}$$

where $A(\xi)$ is also known as the mean force potential or PMF.

All systems but those with very simple partition functions make it extremely difficult to determine the exact phase-space integrals in the previous two equations. The ensemble average is equivalent to the time average for ergodic systems, as described in the section above, because every point in the 6N-dimensional phase space gets sampled during the simulation, guaranteeing unlimited sampling. This can be represented by the following equation:

$$Q(\xi) = P(\xi) = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^{\tau} \rho(\xi(t))dt \tag{2.39}$$

Through Equation (2.39), t stands for time, and ρ "counts" the number of times ξ occurs within a specific time. In MD simulations, the trajectory of the system evolves in time, and by measuring the time-averaged properties of the system, $P(\xi)$, information about the equilibrium properties of the system can be obtained. The free energy, $A(\xi)$ is one such property that can be estimated from time averages.

Our ability to run MD simulations is constrained by the finite amount of computational time we have available. As a result, only a portion of the entire configuration space can be sampled by the time averages we obtain and only the regions around the minimum of $U(\{\mathbf{r}\})$ are sampled quite precisely but the regions with higher energy are less traversed as during the finite time the simulation is run, the system is not able to push itself to those high energy regions which cannot be reached with just thermal energy available and no external support. However, it is important to sample both the rare occurrences that take place in the high-energy regions as well as the areas in configuration space surrounding the minimum in order to derive the potential of mean force (PMF) profile $A(\xi)$ along the selected response coordinate. These rare events are crucial for determining the barriers and transitions between different states in the system. To sample these events multiple techniques have been developed and most of those can be categorized into following:

- **Equilibrium sampling methods**: These techniques require simulating the system under multiple thermodynamic circumstances that span the pertinent region of configuration space.
- **Nonequilibrium sampling techniques**: These techniques entail causing the system to deviate from equilibrium, usually by imposing a biassing potential.
- **Methods with additional degrees of freedom**: These techniques entail adding extra variables that are associated to the desired reaction coordinate and then calculating the free energy while taking into account the impact of the additional degree(s) of freedom.

2.6.1 Accelerated Sampling Techniques Based on Equilibrium Properties

Equilibrium has a property called the potential of mean force (PMF), which depicts the difference in free energy between two states along the reaction coordinate. This indicates that it is solely dependent on the system's equilibrium distribution and not on the process by which the system reached that distribution. One way to determine the free-energy difference in molecular dynamics simulations is to sample the system while it is in equilibrium and compute the PMF along the selected reaction coordinate to discover the transition barrier. An alternative approach is to modify the energy expression in order to lower the energy barrier. The energy function can be modified to reduce the free energy barrier between two states by including a biassing potential. A popular technique for doing this is known as "umbrella sampling" [65, 116] enhanced sampling is accomplished by adding an external biassing potential along the selected reaction coordinate. By flattening the free-energy surface along the reaction coordinate with the biassing potential, regions that are generally difficult to sample in traditional MD simulations can be better sampled. Combining data from many biassed simulations and using techniques like WHAM [117, 118] or umbrella integration [119] can be used to calculate the free-energy profile along the reaction coordinate.

Despite the fact that the free-energy difference is an equilibrium property, it can be calculated using either equilibrium or nonequilibrium methods. Particularly, nonequilibrium methods require much more computational work to simulate the system over much longer timescales than equilibrium methods do. Additionally, rigorous validation is necessary for nonequilibrium methods to make sure the selected path is, in fact, a true transition pathway.

2.6.2 The Umbrella Sampling Method

A computational method called umbrella sampling [65, 116] is used to sample rare events like molecular changes between states. When using umbrella sampling, which was initially introduced by Torrie and Valleau in 1977, the system's Hamiltonian is constrained along a particular reaction coordinate, also known as a collective variable. Targeted sampling of areas within the free energy landscape that would otherwise be difficult to study is made possible thanks to the intentional biassing of the system. The addition of the bias potential can be represented by the following equation:

$$U^{b}(\mathbf{r}) = U^{u}(\mathbf{r}) + w_{i}(\xi) \tag{2.40}$$

Here, w_i is the bias potential of i^th window and is a function of only reaction coordinate, ξ. And superscript b and u represent biased and unbiased terms respectively. Looking back to Equation (2.37) we can find the unbiased distribution, which helps in the calculation of unbiased free energy, $A_i(\xi)$. The distribution can be represented as:

$$P_i^{u}(\xi) = \frac{\int \delta[\xi(\{\mathbf{r}\}) - \xi] \ \exp\{-\beta U(\{\mathbf{r}\})\} \, d^{N}\mathbf{r}}{\int \exp\{-\beta U(\{\mathbf{r}\})\} \, d^{N}\mathbf{r}} \tag{2.41}$$

Similarly, biased probability distribution along a certain reaction coordinate, ξ, produced by a biased system MD simulation can be represented as the following by using Equation (2.40) and Equation (2.41):

$$P_i^b(\xi) = \frac{\int \delta[\xi(\{\mathbf{r}\}) - \xi] \; \exp\{-\beta[U(\{\mathbf{r}\}) + w_i(\xi)]\} \, d^N\mathbf{r}}{\int \exp\{-\beta[U(\{\mathbf{r}\}) + w_i(\xi)]\} \, d^N\mathbf{r}} \tag{2.42}$$

Because all degrees of freedom except ξ are included in the integration in the numerator, the aforementioned equation can be further simplified. The bias only depends on ξ and by using this we get:

$$P_i^b(\xi) = \exp\{-\beta w_i(\xi)\} \times \frac{\int \exp\{-\beta U(\{\mathbf{r}\})\} \, \delta[\xi(r) - \xi] \; d^N\mathbf{r}}{\int \exp\{-\beta[U(\{\mathbf{r}\}) + w_i(\xi)]\} \, d^N\mathbf{r}} \tag{2.43}$$

From the previous two equations, we can represent $P_i^u(\xi)$ as:

$$\begin{aligned} P_i^u(\xi) &= \exp\{\beta w_i(\xi)\} \times P_i^b(\xi) \times \frac{\int \exp\{-\beta[U(\{\mathbf{r}\}) + w_i(\xi)]\} \, d^N\mathbf{r}}{\int \exp\{-\beta U(\{\mathbf{r}\})\} \, d^N\mathbf{r}} \\ &= \exp\{\beta w_i(\xi)\} \times P_i^b(\xi) \times \frac{\int \exp\{-\beta U(\{\mathbf{r}\})\} \; \exp\{-\beta w_i(\xi)\} \, d^N\mathbf{r}}{\int \exp\{-\beta U(\{\mathbf{r}\})\} \, d^N\mathbf{r}} \\ &= \exp\{\beta w_i(\xi)\} \times P_i^b(\xi) \times \langle \exp\{-\beta w_i(\xi)\} \rangle \end{aligned} \tag{2.44}$$

Using this equation we can write free energy, $A(\xi)$ as:

$$A_i(\xi) = -(1/\beta) \ln P_i^b(\xi) - w_i(\xi) + C_i \tag{2.45}$$

where, a biased MD simulation yields $P_i^b(\xi)$, an additive constant $C_i = -(1/\beta) \ln \langle \exp\{-\beta w_i(\xi)\} \rangle$ does not depend on ξ and finally $w_i(\xi)$ is provided analytically. It is important that at least one window spans the entire range of ξ to be studied for Equation (2.45) to output accurate results with C_i being arbitrarily defined. But if that is not the case C_i needs to be calculated using below mentioned equation for each window in the desired range of ξ:

$$\begin{aligned} \exp\{-\beta C_i\} &= \langle \exp\{-\beta w_i(\xi)\} \rangle \\ &= \int P^u(\xi) \; \exp\{-\beta w_i(\xi)\} \, d\xi \\ &= \int \exp\{-\beta[A(\xi) + w_i(\xi)]\} d\xi \end{aligned} \tag{2.46}$$

2.6.3 Harmonic Bias Potentials

The bias potential should ideally be selected so that it is strong enough to promote sampling of all pertinent configurations within the window but not so strong as to introduce large perturbations that could influence the results. Utilizing a harmonic bias potential that is based on how far the coordinate is from a desired value, such as the window's centre, is an useful tactic. The bias potential usually depends on the force constant that controls the curvature of the potential energy surface.

The range of the reaction coordinate ξ is frequently split into numerous "windows" or "bins" to permit sampling across various regions of space. This is due to the difficulty in overcoming rugged energy landscapes and energy barriers that frequently accompany transitions between various sections of the system's free energy surface. To overcome these barriers, bias potential is added to each window that reduces the height of the energy barriers within that window, making it easier to explore that region of ξ space.

The bias potential for every window i can be defined by the following equation:

$$w_i(\xi) = \frac{K}{2}(\xi - \xi_i^{ref})^2 \tag{2.47}$$

Here, K is the force constant that determines the strength of the bias potential, and ξ_i^{ref} is the desired value of ξ that the system should sample in the current window. This bias potential creates a harmonic potential well centered on ξ_i^{ref} that attracts the system towards this value of ξ.

In umbrella sampling, the choice of the force constant K in the harmonic bias potential is extremely important. How strongly the system is drawn to the target value of the reaction coordinate in each window depends on the bias potential's strength. The bias potential might not be strong enough to successfully sample the window if the force constant is too weak, which would produce subpar statistics and inaccurate results. On the other hand, the bias potential may significantly perturb the system if the force constant is too strong, which could lead to errors in the free energy estimate as is typically the case in analysis methods like WHAM [117, 118]. To ensure sufficient overlap between windows, care should be taken in selecting the size and number of windows in addition to the reaction coordinate. The WHAM [117, 118] analysis, which combines the probability distributions from different windows to estimate the free energy surface and determines the accuracy of the last free energy estimate, depends on how much overlap there is. The estimation of the free energy surface is better the more overlap there is between the windows.

To produce accurate and trustworthy results, it is also crucial to sample the phase space as thoroughly as possible in each window. Sampling should be optimised to ensure that the probability density of the system is accurately estimated in each window. This can be done by selecting the reaction coordinate, ξ carefully. It should be a reliable collective variable that accurately captures the system's relevant physics.

2.6.4 Weighted Histogram Analysis Method(WHAM)

By using a harmonic bias potential to bias the system's potential energy, umbrella sampling improves sampling of the reaction coordinate space. However, it is challenging to directly calculate the PMF from the sampled configurations due to the bias potential. This is where WHAM comes in, it is a method for combining the probability distributions obtained from each window to obtain an unbiased estimate of the PMF. To do this, WHAM adds a set of weights, $\wp_i(\xi)$ that take into account the bias created by each window's harmonic bias potential and can be represented by the following equation:

$$P^u(\xi) = \sum_i^{\text{windows}} \wp_i(\xi) P_i^u(\xi) \tag{2.48}$$

The probability distributions obtained from each window, $P_i^u(\xi)$, are given these weights in order to more closely resemble the system's unbiased probability distribution, with the goal of minimising the statistical error of $P^u(\xi)$, which leads us to the following:

$$\frac{\partial \sigma^2(P^u)}{\partial \wp_i} = 0, given \sum \wp_i = 1 \tag{2.49}$$

As summation of all weights is 1, we get:

$$\wp_i = \frac{a_i}{\sum_j a_j}, \quad a_i(\xi) = N_i \exp\{-\beta w_i(\xi) + \beta C_i\} \tag{2.50}$$

N_i represents the total amount of steps sampled for window i. Equation (2.46) is used to calculate C_i.

$$\exp\{-\beta C_i\} = \int P^u(\xi) \exp\{-\beta w_i(\xi)\} d\xi \tag{2.51}$$

Iterating these until convergence is necessary because $P^u(\xi)$ is plugged in Equation (2.51) and C_i is inserted in Equation (2.48) using Equation (2.50). This convergence may take some time for large numbers of bins N_i, so an optimal value of bins should be taken as the tradeoff is between accuracy and computaional time.

2.6.5 Efficient Simulation Techniques for Molecular Dynamics: Periodic Boundary Conditions, Truncation, and Minimum Image Convention

Periodic boundary conditions are used in the molecular dynamics simulations used in this study to determine the bulk and thermodynamic properties of the system. These simulations use periodic boundary conditions as an example of a boundary condition to simulate a system that repeats indefinitely. In a MD simulation, forces like electrostatic or van der Waals interactions are used to describe how the system's particles interact with one another. It takes simulating a lot of particles over a long period of time to accurately represent these interactions. However, it is neither computationally feasible nor practical to simulate an infinitely large system. Instead, periodic boundary conditions represent a finite system that behaves as though it were repeating in all directions as shown in Figure 2.4 for a 2D periodic boundary condition example. This is accomplished by treating the simulation box's edges as though they were connected, causing a particle to cross one edge to reappear at the opposite edge. Periodic boundary conditions effectively eliminate the effects of the boundaries on the interactions between the particles, allowing simulation of a much larger system than would be feasible otherwise. This is crucial for

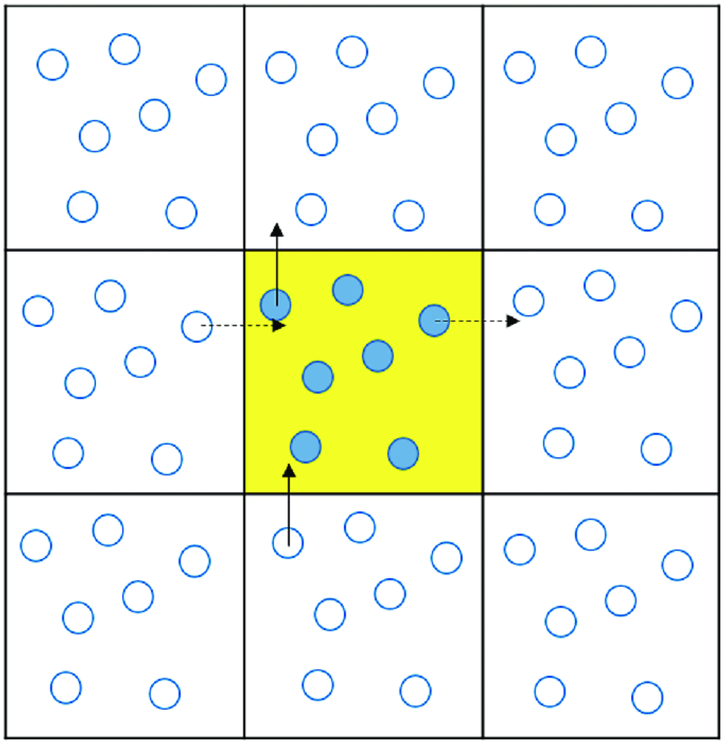


Figure 2.4: An image particle enters the simulation box to take the place of the particle as it exits. True and image neighbours are both taken into account in the computation of particle interactions inside the cutoff range. Illustration and caption taken from [120].

accurately capturing the behaviour of systems in the condensed phase, where long-range and significant particle interactions take place.

When using periodic boundary conditions, it is crucial to take into consideration the range of the system's interactions. The cutoff radius, or maximum distance at which interactions between particles are taken into account, is a common way to express the range of interactions between particles in molecular dynamics simulations. If the range of interactions is broad in comparision to the size of the simulation box, the periodic boundary conditions may have a significant impact on how the system behaves. On the other hand, the periodic boundary conditions are unlikely to have a significant effect on the behaviour of the system if the range of the interactions is much smaller than the size of the simulation box.

In MD, non-bonded interactions like van der Waals, which typically have a long range and can affect particles distant from one another, can be numerous and complicated. The computation of these interactions for each particle in a large system can be very time and resource intensive. Because non-bonded interactions are generally weak and only matter when atoms are close to one another. Because the interaction energy is proportional to the inverse sixth power of the interatomic distance, the strength of the interaction rapidly decreases as the interatomic distance increases. Therefore, a non-bonded cutoff is typically used to prevent measuring the majority of non-bonded pairwise interactions, thereby reducing the number of unnecessary pair-wise energy data produced at each step.

In molecular dynamics simulations with PBC, the minimum image convention is a technique used to handle interactions between particles that are separated by a distance greater than the size of the simulation box. The fundamental premise of the minimum image convention is that each particle in a simulation box is surrounded by an infinite number of periodic repeating images of itself due to PBC. Only the image of the particle closest to the interacting particle is considered when calculating the interactions between particles. The "minimum image" is the one that is closest. To accomplish this a cutoff value is set. By making all atom pair interactions that are farther away than the cutoff value to 0, the number of calculations required for the simulation is significantly reduced, which makes the simulation more computationally efficient.

Molecular dynamics simulations done in this study would benefit from maintaining separate cutoffs for the two types of interactions (van der Waals and Coulomb interatyions) because electrostatic interactions, like Coulomb interactions, have a much longer range than van der Waals interactions. Additionally, the use of highly charged species like DNA in this study emphasises the need for distinct cutoffs.

Even though a cutoff can help to reduce the number of interactions that must be calculated, it still necessitates checking the distance between all pairs of particles to determine which interactions should be included and which should be excluded, so the efficiency of computing the non-bonded interactions may not be significantly improved by the use of a cutoff alone. Therefore, this operation may still be computationally expensive, particularly in large particle systems.

When computing a MD simulation's non-bonded interactions, the use of a cutoff is frequently combined with other methods, such as neighbour lists [114, 121, 122]. The simulation's computational cost is greatly reduced by the use of neighbour lists [114, 121, 122], which effectively identify the particles that are spatially close to one another and only compute the interactions between these particles. A list of nearby atoms that are close enough to potentially interact is stored in an array for each particle in the neighbour list algorithm. To keep it accurate and current, this list of neighbours is updated periodically or whenever a particle moves a certain distance making $\mathrm{O}(\mathrm{N}^2)$ it's worst-case time complexity, where N is the total number of simulation particles. This indicates that the algorithm's computational cost increases quadratically as the number of particles increases. In this worst-case scenario, every simulation particle interacts with every other simulation particle, necessitating the rebuilding of the neighbour list at each time step. For large systems with numerous particles, this can become unaffordable because of the exponentially increasing number of interactions that must be computed.

To overcome the aforementioned obstacle, cell lists algorithm which is a variation of the neighbor list algorithm is used. It segments the simulation box into more manageable cells or bins. An interaction list is created specifically for particles in the same cell or nearby cells using this method. Based on

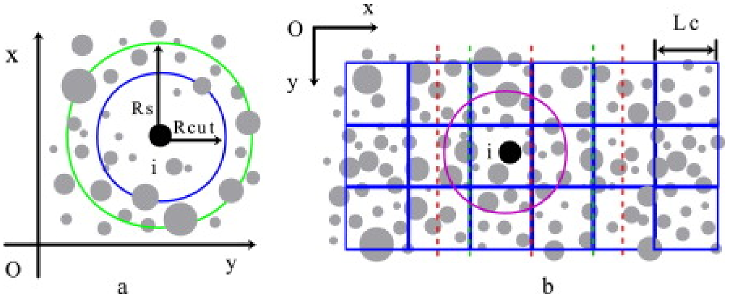


Figure 2.5: (a) All particles lying in R_s radius are included in the neighbour list for every atom i, (b) Cell list optimisation where every cell is of length L_c, Illustration taken from [123]

where it is located within the simulation box, each particle is given a cell to belong to. The worst-case time complexity of the cell list approach is $O(N)$, where N is the total number of particles in the simulation. In this worst-case scenario, each particle in the simulation is given its own cell, necessitating the rebuilding of the cell list after each time step. In large systems with many particles, this can still get expensive, but it is much less taxing than the worst-case scenario for the neighbour list algorithm.

2.6.6 Long-Range Electrostatic Interactions: Ewald Sum and PME

As the electrostatic interactions are long-ranged, which do not deteriorate quickly enough to allow for small cutoffs [124, 125], Ewald sums are required in MD simulations. In a simulation cell with periodic boundary conditions, the number of charged particle images grows in direct proportion to the size of the simulation box. Consequently, it is necessary to compute a large number of interactions.

The Ewald sum is a method for calculating electrostatic interactions in a molecular dynamics simulation that divides the electrostatic potential into two components: long-range and short-range. The long-range component is computed by adding the contributions from an infinite number of charged particle images, while the short-range component is calculated using a real-space cutoff [126, 127].

The Particle Mesh Ewald (PME) method uses a combination of Fourier transform techniques and the Ewald summation method to handle the long-range interactions that occur in charged systems. The PME algorithm divides the system into a grid with a regular mesh, where each mesh point represents the density of an electron cloud [124, 125, 128]. Then, Fourier transforms are used to efficiently calculate the electrostatic interactions between the particles throughout the entire system. This calculation yields a 3D grid that displays the system's electrostatic potential. This allows for a much more efficient computation of the electrostatic interactions, as the long-range interactions are only calculated on the mesh grid, rather than between all pairs of atoms in the system, which makes the overall calculation much faster.

2.6.7 Bond Length Constraints and their Implementation

The interval of time between two updates of the particle positions and velocities is known as a timestep, Δt. In molecular dynamics simulations, the selection of Δt is crucial as it affects the simulation's stability, the precision of the results, and the computation time. The timestep must be small enough for the numerical integration of the motion equations to accurately represent the system's behaviour, but if it is too small, the simulation will be very slow because many steps will be required to achieve the desired simulation time. And if the timestep is too large, the simulation may become unstable, as the integration method may no longer accurately describe the evolution of the system over time.

Intramolecular bonds have extremely high vibration frequencies [59], typically in the terahertz range or higher. This means that bond vibrations happen on much faster time scales than the time steps used in classical molecular dynamics simulations, which are typically on the order of femtoseconds. To accurately represent these fast bond vibrations, extremely small time steps would be required, making the simulation very computationally expensive. Instead of attempting to represent the bond vibrations directly, it is common practise in classical computer simulations to use constraint algorithms, such as SHAKE or RATTLE [129–131], to keep the intramolecular bonds at their ideal lengths. The intramolecular bonds, particularly hydrogen bonds, have extremely high vibration frequencies, as was already mentioned. These high frequencies may cause numerical instability in the simulation if the time step is too large. The high-frequency vibrations are effectively constrained and the time step can be safely increased (2 fs in this study) leading to a more effective simulation by removing the motion of hydrogen atoms through the aforementioned algorithms. This improved efficiency enables longer simulation runs while also lowering the overall computational cost of the simulation. The SHAKE algorithm implemented in AMBER is used for all MD simulations in this study.

Chapter 3

CPD lesion recognition by the DNA Damage Sensing Protein Rad4/XPC: Energetics and Mechanism

3.1 Introduction

DNA plays a central role in the life of a biological cell. DNA damage can give rise to deleterious cellular responses and cellular malfunction, which can lead to eventual cell death or uncontrolled cell growth[69, 132–136]. DNA repair proteins can sense and repair DNA damage to safeguard the genome integrity of cells. The UV light-induced cyclobutane pyrimidine dimer (CPD) is the most common type of UV-induced DNA damage. This form of DNA damage has been linked to a variety of skin-related genetic diseases in humans [48, 66, 137, 138].

The pathway known as nucleotide excision repair (NER) plays a crucial role in repairing cyclobutane pyrimidine dimers (CPDs), it constitutes the detection of structural distortions in DNA that contains CPD by a particular repair protein, followed by the mobilization of additional proteins to repair the DNA damage [139–147]. It appears that damage verification and further NER activities need the repair protein's pausing and subsequent conformational changes at the lesion site after it slides corkscrew-like along the DNA to scan and locate for the lesion [148–153]. A crucial protein called Xeroderma pigmentosum C (XPC) helps mammalian cells repair cyclobutane pyrimidine dimers (CPDs) [144–147]. The ATP-dependent helicases (XPB and XPD) in transcription factor II H (TFIIH) unwind the DNA duplex after XPC locates the site of damage, creating a bubble surrounding the lesion [154–166]. To restore the complete DNA structure, the lesion-containing oligonucleotide is then removed by endonucleases (XPG and XPF), the resulting gap is filled with DNA polymerase, and it is finally sealed with a DNA ligase [167–172].

For research on XPC-mediated DNA damage repair in human beings, studying Radiation Sensitive 4 (Rad4), the yeast equivalent of XPC, can be helpful. Rad4 can be used as a valuable model for studies in this area because it shares several structural and functional similarities with XPC [78, 86, 87]. The CPD containing DNA fragments linked to Rad4 in its crystal structure offers a valuable starting point for research into the molecular mechanisms of Rad4-mediated DNA damage repair in yeast cells. This structure acts as a fundamental model that may be consulted in order to comprehend the corresponding XPC-related human processes.[83].

The protein, Rad4 consists of 3 β-hairpin domains (BHD1, BHD2, and BHD3) along with an N-terminal transglutaminase domain (TGD). [85, 86]. The BHD1 domain and the transglutaminase domain (TGD) are crucial for preserving the undamaged DNA segment's structural integrity. These domains bind to the undamaged DNA and aid in maintaining its integrity. The β-hairpin of BHD3 fills the gap created by the CPD and its adjacent bases that have been flipped out by inserting itself into the DNA major groove during the repair process. BHD2's β-hairpin interacts with the minor groove of the DNA around the lesion while also forming hydrogen bonds with the DNA's backbone. The BHD2-BHD3 binding interface holds these ejected partner bases.

Base-base hydrogen bonding patterns are changed by CPD, which also has an impact on the DNA's base pairing stability around the damage site. Because of this distortion of the DNA's general structure, the DNA at the lesion location is bent and unwound. For instance, in the absence of Rad4, the CPD-containing DNA is approximately 30° bent towards the primary groove and approximately 9° unwinding towards the lesion site [67, 173, 174]. These small structural distortions in DNA, which are produced by the inability of CPD to make hydrogen bonds with its companion bases, serve as a signal for damage identification by repair proteins [69, 138, 175–177]. The DNA's bending angle increases to around 42° when Rad4 attaches to the damaged DNA, and there is considerable helical unwinding close to the active site [86].

Although providing a static perspective of damage identification by Rad4 through the aforementioned structural insights acquired from the Rad4-DNA crystal structure, the overall process is probably rather dynamic. Unsurprisingly, the intricacy of multiple dynamic events that take place during damage identification and repair by Rad4 cannot be effectively accounted for by the structure-based static view. Previous studies have provided insight into the key initial processes involved in Rad4's ability to detect lesions. Included in these are Rad4's association with DNA, the flipping out of the partner bases and the lesion from the DNA duplex, as well as the insertion of β-hairpins into the grooves of the DNA. Each of these occurrences is connected with a certain mechanism and time frame. For instance, the Rad4-DNA connection is anticipated to start with TGD and BHD1 binding to the portion of DNA that is not damaged, then BHD2 and BHD3 attaching to the portion of DNA that has the lesion. The CPD lesion and its companion thymines should then completely flip out of the DNA duplex as a result of the BHD3-β-hairpin and BHD2-β-hairpin inserting into the main and minor grooves of the DNA, respectively. It appears that the connection of Rad4 with DNA must have occurred before the flipping of the partner bases since the partner bases could not flip out of the CPD-containing DNA duplex in the absence of Rad4 [67]. The sequence in which the 5'-dA and 3'-dA partner bases flip, as well as whether or not these flipping events are successful in inserting the BHD3 β-hairpin into the lesion location, are yet unknown. The likelihood of these molecular processes cooperating and correlating with one an-

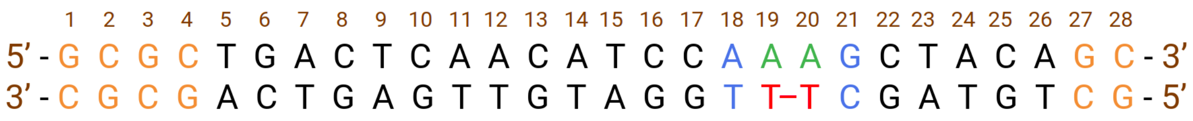

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
5'-G C G C T G A C T C A A C A T C C A A A G C T A C A G C-3'
3'-C G C G A C T G A G T T G T A G G T T–T C G A T G T C G-5'

Figure 3.1: **DNA sequence**: The nucleotides in red represents the lesion, CPD.

other presents another difficulty. That is, the NER process may fail if any of these things went wrong. Determining the exact sequence of these processes, as well as their molecular mechanics, energetics, and interconnectedness at the atomic level, is crucial. The current work uses molecular dynamics and improved sampling simulations to examine the processes, energies, and sequence of these events as well as any potential linkages between them.

3.2 Materials and Methods

3.2.1 Models

3.2.1.1 Pre-association Encounter Complex

The pre-association encounter complex represents a state of the system prior to the association of Rad4 and the damaged DNA. It can be thought of as an intermediate state between the dissociated reactant state and the final bound complex. Due to unavailability of the crystal structure of this state, we first built a model of this state using the following protocol: a canonical B-DNA of the desired sequence (Figure 3.1) was built using the Nucleic Acid Builder(NAB). A major part of this sequence was taken from the crystal structure of a damaged DNA [84], while the rest were added (Orange coloured nucleotides in Figure 3.1). A cyclobutane pyrimidine dimer (CPD) lesion produced from an unbound CPD-containing DNA crystal structure (PDB ID: 1T4I) [67, 68] was used to replace two consecutive thymines in our modelled DNA structure. This model of CPD-containing DNA will henceforth be referred to as the *lesioned DNA*. Also, the 19th and 20th nucleotide on the undamaged strand (Nucleotides in green colour in Figure 3.1) will be attributed as 5'-dA and 3'-dA respectively.

The crystal structure of DNA-free apo-Rad4 (PDB ID: 2QSF) was used to model the unbound Rad4. The missing residues of apo-Rad4 were modelled using the protein-Modeller [86, 179]. Given the models of the unbound lesioned DNA and apo-Rad4, we proceeded to dock them suitably to build the pre-association encounter complex. The docking was performed in two stages; firstly, the unbound lesion-containing DNA was aligned with the Rad4-bound DNA of the open complex (PDB ID: 2QSG). Secondly, TGD and BHD1 of the apo-Rad4 was aligned with the DNA-bound Rad4 of the open complex. This state in which both the unbound lesioned DNA and apo-Rad4 are aligned to achieve optimal

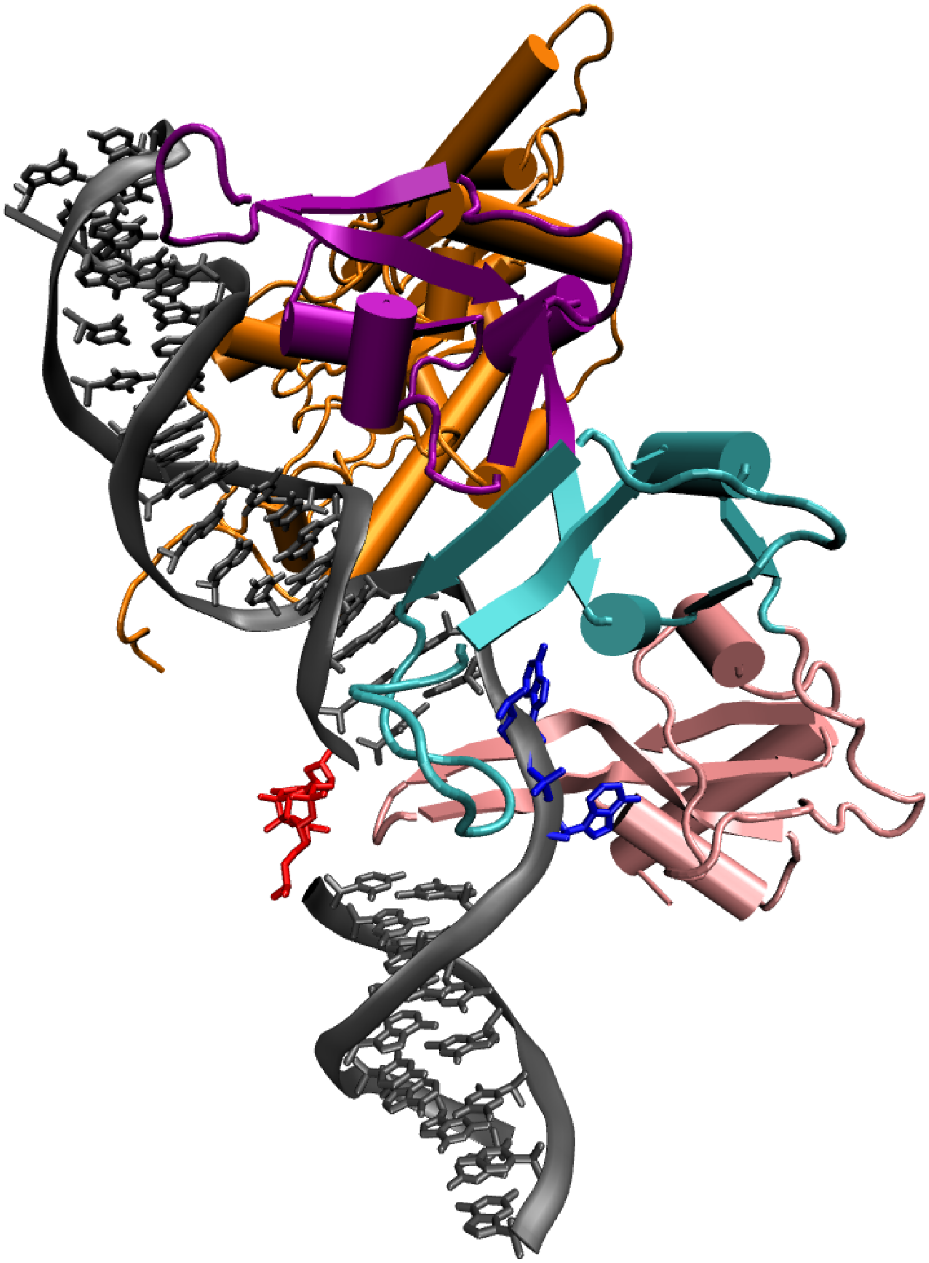


Figure 3.2: Model of CPD-containing DNA-Rad4 post-recognition complex (PRC). Color code: TGD (orange), BHD1 (purple), BHD2 (cyan), BHD3 (pink) domains of RAD4 and the CPD (red) and its partner adenine bases (blue) of DNA (grey). The image was generated using VMD [178].

overlap with the final open complex is used as a model of the pre-association encounter complex in the present study.

3.2.1.2 Associated Open Complex

We used the crystal structure of the Rad4-DNA complex (PDB ID: 2QSG) as the basis for our model to generate the final associated open complex. The β-hairpins of Rad4's BHD2 and BHD3 are located close to the lesion in the minor and major grooves of the DNA in this bound form. Importantly, in this structure, the CPD and the adenine bases that it is accompanied by on the undamaged strand are fully displaced or ejected from the DNA duplex. A CPD-lesion is found at the 19^{th} and 20^{th} base pairs of the 28-base pair DNA sequence (See Figure 3.1) that was extended from this crystal structure. The partner bases A19, $A20_{\mathrm{u}}$ of the CPD on the undamaged strand (designated by a subscript u) shall be referred to as 5'-dA and 3'-dA, respectively, in the text that follows. The same approach that was previously utilised to construct the pre-association encounter complex was employed to add a CPD lesion to the open complex since the coordinates of the CPD lesion were not resolved in the crystal structure of the open complex. Additionally, we modified each of the two mismatched thymine partner bases discovered opposite the lesion in the crystal structure of the open complex using the Swapna module of UCSF Chimaera [180, 181]. We specifically substituted adenine bases for these mismatched thymines. The main goal of selecting this precisely matched lesion-containing DNA is to explore the single roles played by the lesion in this process and to remove the influence of mismatch on Rad4-DNA binding. Figure 3.2 depicts the model of the related open Rad4 and DNA complex that forms.

3.2.1.3 Intermediates of Rad4-DNA Complex

The structural models of the important intermediate states of the DNA-Rad4 complex that include CPD (models B, C, D, E, and F in Figure 3.4) were thought to be a good way to investigate the processes in between damage recognition and repair by Rad4/XPC. Model B represents an intermediate stage in which the BHD3-β hairpin is removed from the Rad4-DNA complex's final bound state. Similar to model B, models C and D have one of the CPD partner bases (5'-dA or 3'-dA) flipped into the DNA duplex. Model E is similar to Model B but flips both partner bases into the DNA duplex. CPD lies outside of the DNA duplex in models C through E. Identical to Model E, Model F essentially flips the CPD into the DNA duplex. In the broader NER process, each of these models reflects the conclusion of a believable intervening phase. In this case, a transition between two such models corresponds to an interesting intervening process. For instance, the deinsertion (model A → model B) or insertion (model B → model A) of the BHD3 β-hairpin from the DNA duplex are examples of transitions between models A and B. One may think of the whole NER process as a series of these in-between transitions between model A and model F. We modelled several event sequences that corresponded to various conceivable processes for the NER process, as illustrated in Figure 3.4.

We looked into numerous potential mechanisms for the NER process by analysing various intervening processes between intermediate states utilising relevant collective variables (see later). Multiple unbiased NPT molecular dynamics (MD) simulations were run starting from structures with varying values of the collective variable (CV) unique to each intervening process in order to model these intermediate states. When the CV time series from these simulations was evaluated, it was discovered that a sizable fraction of trajectories converged around a certain value of the CV. The structures derived from these trajectories with the CV value near to this converged value were then grouped with regard to important nucleotides ($C17_u$ - $C22_u$ and $G17_d$ - $G22_d$) surrounding the damaged DNA lesion location using the root mean square deviation (RMSD) approach. For instance, BHD3 β hairpin was de-inserted to various degrees in order to produce model B, and all of those de-inserted states were utilised in the objective manufacturing run. Then, based on their RMSD values of the active site with every other conformation in the pool, all the configurations/structures formed after certain intervals during all those production runs were taken in a cluster pool and separated into several clusters. The intermediate state was determined to be the cluster centroid of the biggest of the ten clusters that were generated. The biggest cluster ($C17_u$ - $C22_u$ and $G17_d$ - $G22_d$) shows the conformation of the active site that is most likely to exist since it contains the majority of conformations with minor variations.

3.2.2 Molecular dynamics simulation

To conduct all-atom molecular dynamics simulations of the model systems, we utilized the AMBER 2018.10 simulation software [182, 183]. We used the ParmBSC1 force field [184] for the DNA components and the ff14SB force field [107] for the protein components. In the simulation, water molecules were represented using the TIP3P model [185]. We used the Antechamber module of AmberTools19 to give the atoms in the CPD lesion partial charges. The general Amber force field (GAFF) [186] was used to determine the lesion's residual force field parameters. This method made sure that the CPD lesion in the molecular dynamics simulations was properly characterised. In a TIP3P water box with 20Å of water padding in all directions, each of these model complexes was solvated. In the molecular dynamics simulations, the SHAKE algorithm [129] was employed to constrain the lengths of bonds containing hydrogen atoms. In every dimension, periodic boundary conditions were used. The simulations used a 10Å direct space cutoff and a 10^{-5} tolerance. Particle mesh Ewald (PME) [128] was used to handle long-range electrostatic interactions, with a 4^{th} order B-spline interpolation and an Ewald coefficient of 0.27511. With a 10Å limit, Van der Waals interactions were also taken into consideration. The electrostatic and van der Waals interactions were appropriately handled in the simulations due to these settings.

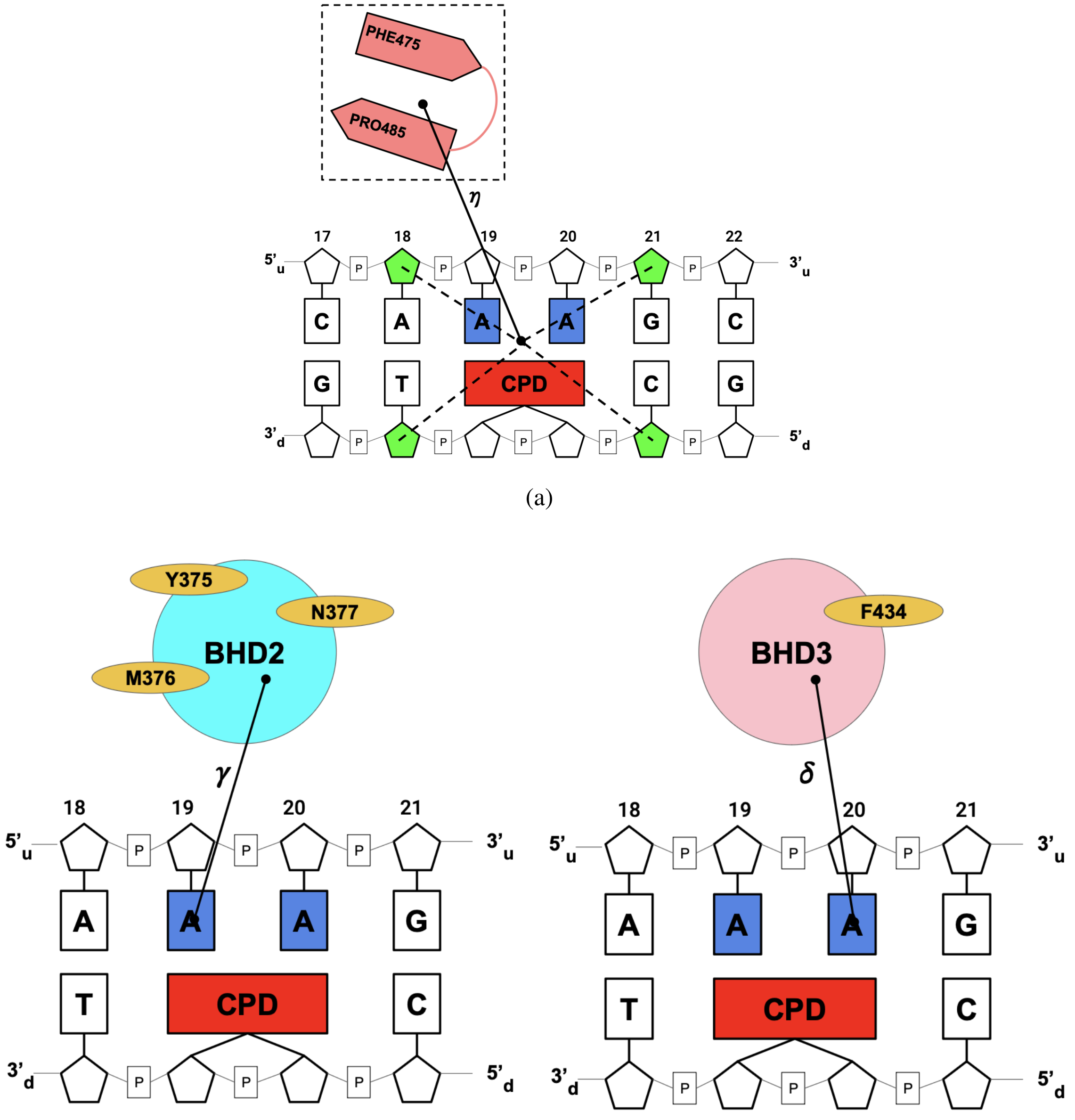
Strong harmonic constraints (spring constant of 100 kcal mol$^-$1Å$^-$2) were applied to the crystallographically resolved atoms of the complex to hold them close to their experimentally resolved positions, while weak harmonic constraints (spring constant of 10 kcal mol$^-$1Å$^-$2) were applied to the unresolved atoms of the complex whose coordinates were conjectured in order to maintain the overall structural integrity of the DNA-Rad4 complex during the initial phase of energy minimization. The harmonic restrictions on the unresolved atoms were eliminated in the following step of energy minimization, but those on the resolved atoms were kept. At every level of energy minimization, the water molecules and counter ions were unrestrained. The energy-minimized configurations were then equilibrated for 0.02 ns in the NVT ensemble at 300 K, followed by 2 ns simulations in the NPT ensemble at 300 K and 1 bar, while maintaining the harmonic restrictions on the resolved atoms. After removing all constraints, each system was subjected to energy minimization and equilibration. The entire system underwent energy minimization, followed by equilibration in the NVT ensemble for 0.02 ns and then in the NPT ensemble for 2 ns. The energy minimization convergence tolerance was set at 10^{-4} kcal mol^{-1} Å^{-1}. For each energy minimization run, the steepest descent approach was used for the first 20,000 steps, and the conjugate gradient method for the following 20,000 steps [182]. A Berendsen barostat [115] with a pressure relaxation time of 1 picosecond (ps) was used to keep the pressure at 1 bar. Using a Langevin thermostat [187] with a collision frequency of 1 ps^{-1}, the temperature was kept at 300 K. The velocity Verlet technique [187] was used to integrate the motion equations, with a time step of 2 femtoseconds (fs). With the help of these variables, the system was able to reach a stable equilibrium condition for additional analysis.

3.2.2.1 Umbrella Sampling

We used the umbrella sampling method to capture pertinent conformational changes in the Rad4-DNA complex that take place during NER and to quantify the associated energetics because the timescales of key molecular events of NER are likely to be longer than the accessible timescales of conventional MD simulations. The collective variables for β-hairpin insertion, flipping of partner bases, and Rad4-DNA association are defined in the sections that follow.

3.2.3 Collective variable for β-hairpin insertion:

A distance-based CV, η, was used to assess the insertion of the β-hairpin of the BHD3 into the lesion-site. The distance η between the centres of mass (COM) of all the residues in the β-hairpin of BHD3 and the COM of the nearby bases of CPD and its partners (A18$_u$, G21$_u$, T18$_d$, C21$_d$) (Figure 3.3a). For the equilibration and production runs, the biassing harmonic force constants were adjusted to 75 kcal mol$^-$1Å^{-2} and 5 kcal mol$^-$1Å^{-2}, respectively. For the umbrella sample, ins was adjusted in stages



(a)

(b)

(c)

Figure 3.3: **Schematic representation of the Collective Variables used to simulate the 2 key processes of NER.** (a) The distance denoted by "η " refers to the separation between the center of mass (COM) of the backbone heavy atoms of the BHD3-β hairpin (in pink) and the COM of the sugar rings of the neighboring bases, including the cyclobutane pyrimidine dimer (CPD) and its partner bases, i.e. (A18$_\mathrm{u}$, G21$_\mathrm{u}$, T18$_\mathrm{d}$, C21$_\mathrm{d}$) (green). (b) "γ " refers to the separation between the Adenine base, A19$_\mathrm{u}$ (blue) and the heavy atoms' centres of mass (COMs), which are shown as yellow ellipses in the BHD2 domain pocket. (c) "δ " refers to the separation between the Adenine base, A20$_\mathrm{u}$ (blue) and the centre of mass (COM) of the heavy atoms in the BHD3 domain-belonging amino acid PHE434 (shown as a yellow ellipse).

of 0.5Å from 1Å to 22.5Å. The computed ins value for the Rad4-DNA complex experimental crystal structure (model A) is 4.95Å.

3.2.4 Collective variable for base flipping

A distance-based CV was created independently for each partner base in order to capture the flipping dynamics of the undamaged strand of DNA's partner bases for CPD, 3'-dA and 5'-dA. The partner bases (3'-dA and 5'-dA) are both ejected from the DNA duplex and strongly attached to the binding pocket at the interface between Rad4's BHD2 and BHD3 domains, according to the crystal structure of the Rad4-DNA complex. TYR375, MET376 and ASN377 of Rad4's binding pocket residues form effective contacts with 5'-dA in this ejected extra-helical state. Nevertheless, in the intra-helical state, where 5'-dA is aromatically stacked with its neighbouring $A18_u$ base of the DNA, these interactions are not present. And therefore, it is thought that a relevant CV to characterise the flipping dynamics of 5'-dA is the distance between its centre of mass (COM) and the heavy atoms of the binding pocket residues. This CV will now be referred to as γ as seen in Figure 3.3b. The biassing harmonic force constants for the equilibration and production runs were set to 100 kcal $mol^{-}1Å^{-2}$ and 10 kcal $mol^{-}1Å^{-2}$, respectively. γ was changed from 4.0Å to 19.5Å at increments of 0.5Å.

Likewise to this, 3'-dA aromatically stacks with PHE434 of the Rad4 BHD3 domain in the extra-helical state, however this stacking interaction was missing in its intra-helical state, where it stacks with the nearby $G21_u$ base. As a result, the gap between the COMs of 3'-dA and PHE434 was selected as the ideal CV to explain the flipping dynamics of 3'-dA in the current investigation. This CV will be referred to as δ from here on, as seen in Figure 3.3c. The biassing harmonic force constants were adjusted to 100 kcal $mol^{-}1Å^{-2}$ and 10 kcal $mol^{-}1Å^{-2}$ for the equilibration and production runs, respectively, and δ was changed from 2.0Å to 16.5Å in increments of 0.5Å.

3.2.4.1 Umbrella Sampling Protocol

Each of the aforementioned events underwent an independent umbrella sampling simulation. The final frame of unbiased MD simulations served as the starting framework for these simulations. Each umbrella sampling run began with the displacement of the system to the selected window using a harmonic biassing potential with a high spring constant (k_{eq}), bringing the corresponding CV to the window's centre. This was accomplished by doing a biassed NPT equilibration run for 200 ps for each window. This was then followed by a 6 ns production run in the NPT ensemble at 300 K and 1 atm pressure while under the influence of a harmonic biassing potential with a weaker spring constant (k_{prod}) that is significantly less than keq. For various molecular events of interest, different $k_{eq}$and k_{prod}values have been chosen. The same parameters as unbiased MD runs were used in these umbrella sampling simulations, but with an additional restriction on the following distances: (1) The separation of the COMs of bases

A18$_u$ and T18$_d$ when they are restricted at 6.06Å with a 25 kcal mol^{-}1Å^{-2} harmonic bias (2) the distance applying a bias of 25 kcal mol^{-}1Å^{-2} between the COMs of bases G21$_u$ and C21$_d$. These are the bases that the CPD lesion and its companion adenines are close to. We confined the neighbouring bases in their respective crystalline state conformations in all of our umbrella sampling simulations since we expect them to be in their intra-helical states during the whole NER process [83].

3.2.4.2 Order of events

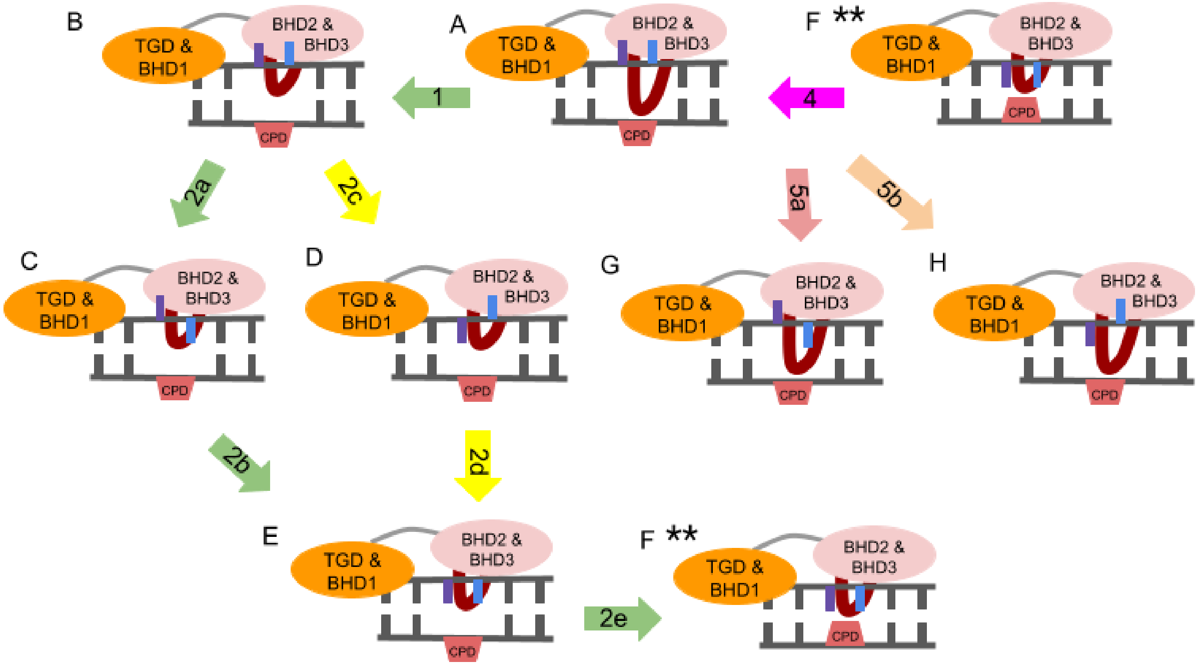


Figure 3.4: Models and sequences of events (denoted by numbered arrows) considered. (A) Rad4-DNA bound complex, (B) bound complex with BHD3 β-hairpin deinserted from the damage site, (C, D) same as (B) except for one of the partner bases (3'-dA (blue) or 5'-dA (violet)) flipped into the DNA duplex, (E) same as (B) but both partner bases are flipped into the DNA duplex, (F**) same as (B) except that both partner bases and the CPD lesion are flipped into the DNA duplex, (G) same as (E) but the BHD2 and BHD3 domains are dissociated from the DNA, (H,I) bound complex with one of the partner bases flipped into the DNA duplex, (J) bound complex with the BHD2 and BHD3 domains dissociated from the DNA. Transitions studied: deinsertion of BHD3 β-hairpin (1); flipping of partner bases (2a, 2b, 2c, 2d) followed by the flipping of CPD (2e) in the β-hairpin deinserted state; Structure marked with * was selected by taking the most probable structure after clustering on 100ns of the unbiased production run of the bound Rad4-DNA complex. Structures marked with ** are the same meta-stable state formed after flipping in CPD.

Using simulations utilising sequential umbrella sampling, the sequence of the aforementioned events and any possible association between them were investigated. According to this method, when an event's umbrella sampling simulation was finished, a meta-stable state structure was made from the produced trajectories and utilised to simulate the next event in the sequence's umbrella sampling.

The association of the BHD2/3 domains with DNA is followed by the flipping out of partner bases, and finally the insertion of the BHD3-β hairpin into DNA, according to past studies on the identification and repair of UV damages in DNA [84, 188]. Therefore, a biassed simulation of Rad4-DNA interaction followed by partner base flipping and insertion of the β hairpin of BHD3 into DNA is necessary to analyse the energetics of these events in this particular order. It is unfortunately difficult to simulate the sequence of events beginning from the pre-associated state since the crystal structure of the Rad4-DNA complex is only known in the post-recognition state, where the aforementioned processes have already taken place. The complicated energy surface of this system, which has multiple pathways leading to different intermediate states and barriers separating them at different heights between the pre-associated state and the post-associated bound complex, makes it even more difficult to understand how the entire process works. We have taken a reverse approach to solving these problems by starting with the experimental crystal structure of the Rad4-DNA complex. By starting with the deinsertion of the BHD3 β-hairpin and then focusing on the flipping in of the partner bases, we concentrate on understanding the reverse process. This method gives us insights into the kinetics and mechanics of these significant repair stages.

To completely understand the flipping behaviour of the partner bases, it is crucial to ascertain whether the 3'-dA and 5'-dA bases flip out simultaneously (in a concerted fashion) or one at a time (in a sequential manner) after Rad4 binds to them. According to earlier studies, the sequential flipping of bases is more advantageous energetically than the concerted mechanism [86, 189, 190]. Yet, there is still much to learn about the energy characteristics of these flipping occurrences as well as whether the 3'-dA base flips before or after the 5'-dA base.

Several umbrella sampling simulations were run utilising various potential flipping mechanisms in order to identify the order of the partner bases' flipping occurrences. The first study concentrated on the 5'-dA base flipping dynamics in two distinct structures, one with an extra-helical 3'-dA base and the other with an intra-helical 3'-dA base. The first model illustrates flipping the 5'-dA base after flipping the 3'-dA base out, whereas the second model shows flipping the 5'-dA base before flipping the 3'-dA base. The BHD2 β-hairpin was partially deinserted during these flipping experiments (Figure 3.4) on the metastable structure of the Rad4-DNA complex, and a comparison of the flipping energy profiles from both experiments can shed light on the partner bases' flipping sequences during Rad4's recognition of DNA damage. Moreover, in the crystal structure of the Rad4-DNA complex, where the BHD3 β-hairpin is precisely inserted into the DNA duplex, the flipping of these bases from their extra-helical locations was examined individually.

We investigated the insertion of the BHD3-β hairpin into the DNA duplex using umbrella sampling simulations. For these simulations, two different Rad4-DNA complex structures were taken into account. In the first structure, the CPD and partner bases were both external to the helical conformation. In contrast, the partner bases and CPD were positioned within the DNA helical conformation in the second structure. We intended to get a thorough understanding of the BHD3-β hairpin insertion process under various circumstances by investigating these two scenarios (Figure 3.4). This analysis was done to determine whether the β-hairpin insertion succeeds or precedes the base flipping. In order to flip the CPD inside the helix, a harmonic biassing potential was applied to the distance between the partner bases' COMs and the CPD's COM. Any of these tests can provide us potential hints about how closely β-hairpin insertion and partner base flipping are coupled.

3.2.4.3 Force Field for CPD

Using the generic Amber force field (GAFF) established in Antechamber, the force field parameters for the CPD lesion were created. Antechamber uses GAFF, which is compatible with the conventional AMBER force fields, to automatically compute charges and atom kinds. The initial structure for running Gaussian from the lesion-PDB (ID: 1SNH) [191] with a net molecular charge of **-2** is first generated using antechamber. Antechamber received the necessary log file from the optimization run in order to produce RESP-based partial charges [192]. The prepgen tool was used to create a mainchain file from the Antechamber output structure file, which would subsequently be utilised by the *parmchk2* programme to check for missing force-field parameters and atoms and atom-types.

3.3 Results and Discussion

The results of umbrella sampling simulations are shown in this section for a variety of molecular events that take place during RAD4-driven identification. These occurrences the flipping of the partner bases of CPD and the insertion of the BHD3 β-hairpin. It's important to remember that these simulations started with the complex's crystal structure in a bound state. The findings of this study reveal the inverted order of events in this identification process as a result. The simulations show how the extra-helical CPD and its partner bases become intra-helical states, as well as how the BHD3 β-hairpin is removed from the DNA double helix.

3.3.1 BHD3 β-hairpin De-insertion

Figure 3.5 depicts the free energy profile, F(η), determined by umbrella sampling and related to the deinsertion of the BHD3 β-hairpin from Model A. The observed crystal structure value of η = 4.94Å is quite close to the single minimum that $F_D(\eta)$ displays at $\eta \sim$ 3Å (denoted by η_{min}). By comparing the energy minimum's location to the crystal structure of the mismatched DNA containing CPD bound to Rad4, a displacement of 2Å is seen. This variation can be attributable to the differences between

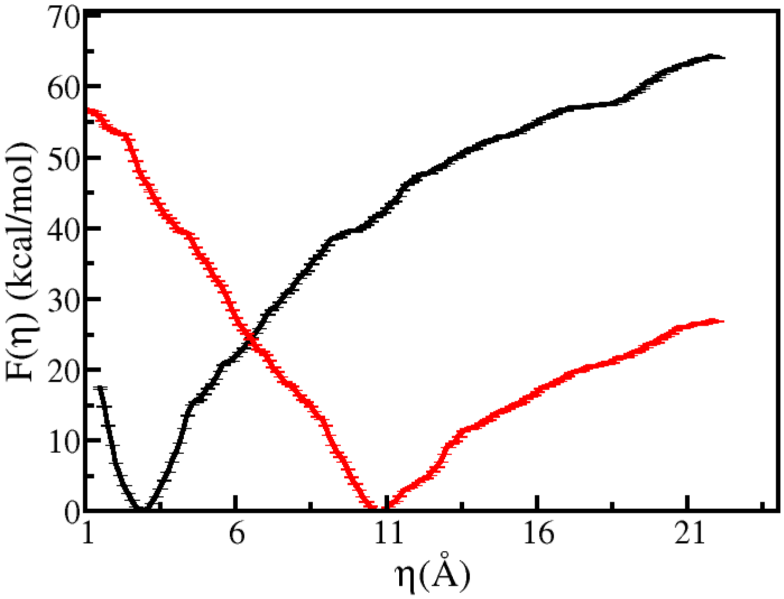


Figure 3.5: The potential of mean force for the deinsertion of the β-hairpin of BHD3 from the lesion site of the damaged DNA duplex for Model A (black) and Model F (red).

the DNA models utilised in our research and the crystal structure. Whereas the crystal structure has a mismatched DNA, our model has a damaged but correctly matched DNA. The CPD in the mismatched crystal structure is also undecidable.For 1.51Å > η > 4.54Å and smaller, the energy basin surrounding the minimum is roughly symmetric, but for η > 4.54Å, there is a clear departure from the harmonic behaviour. The β-hairpin appears to undergo an elastic restoring force due to its favourable interactions with DNA in the harmonic regime (1.51Å > η > 4.54Å) (due to hairpin-DNA van der Waals and polar contacts). By interfering with some beneficial interactions, the β-hairpin appears to rupture this elastic cage at the location of the lesion. When dragged past this harmonic limit, the hairpin displaces away from the damage site. The slope shift in F(η) measured at η = 4.54Å reflects this transition from the harmonic behaviour to the cage-breaking event. The minimum energy required to detach the β-hairpin from the lesion site of DNA in the Rad4-DNA complex is around 13 kcal/mol, which is the difference in free energy between the global minimum and the crossover point at η = 4.54Å. This energy is comparable to the stabilising energy that the BHD3 β-hairpin experiences at the location of the lesion.

The deinserted meta-stable state of the system just prior to the β-hairpin insertion may be ascertained using the β-hairpin deinsertion trajectories produced by the umbrella sampling simulations. We must simulate this metastable condition for at least two reasons since the experimental structure is not accessible. We may first investigate the mechanism and energetics of the β-hairpin insertion into the DNA duplex by beginning from this stage. Second, it enables us to examine the flipping of partner bases and the β hairpin separately, that is, without regard to the influence of the β hairpin that has been inserted on either of their flipping processes. Also, β hairpin insertion to analyse their coupled behaviour after flipping in bases is studied in a subsequent section. The PMF curve's specific shoulder locations (8Å $\leq \eta \leq$ 13.5Å) were chosen to form the meta stable state. These shoulder positions were selected because they all matched to the de-inserted condition and because PHE 475 and 5'-dA did not interact at any of the shoulder points. These shoulder points lack the Hbond interactions between residues 397(O) and 481(H), which present in minima. Moreover, 3'-dA was stacked with PHE 434. Subsequently, a short, unbiased production run was carried out on the chosen shoulder locations. These runs underwent clustering, and the most likely structure was chosen to symbolise the de-inserted meta-stable state. This structure's ins value is 11.46 Å, which translates into a ΔG value of 40.02 kcal/mol. Partner base flipping was then studied using this structure. In earlier studies, it was shown that the lesion was co-relatedly flipped out with the insertion of BHD3-β hairpin in the active site [86]. The lesion was not detected in a full intra-helicular conformation in the de-inserted state since the process is explored in the reverse direction in this work.

It's not yet apparent if the flipping of the partner bases and CPD occurs before or after the β-hairpin insertion. The partner bases and CPD are inferred to be flipped out when the BHD3 β-hairpin is deinserted or inserted since the calculation of $F(\eta)$ was performed for the experimentally determined crystal structure of the bound Rad4-DNA complex Estimating the energetics of the β-hairpin insertion in the presence of intra-helical partner bases and CPD is interesting. On a β-hairpin-deinserted intermediate structure Figure 3.5 of the Rad4-DNA complex with the CPD and partner bases reversed inside, more umbrella sampling simulations were conducted. In other words, these simulations were run with the presumption that the β-hairpin insertion is successfully completed by the partner bases. Furthermore shown in Figure 3.5 is the insertion free energy profile that was produced for this model. In comparison to the free energy profile derived from the crystal structure, this one is very different. For instance, relative to the crystal structure, the free energy minimum is pushed to a larger value of eta. The intra-helical intermediate structure has an η_{min} of 10.7Å whereas the Model A has an η_{min} of 3Å. This suggests that if the CPD and companion bases are intra-helical, the BHD3 β-hairpin cannot reach the damage location. Hence, it would seem that the β-hairpin prefers to remain in the de-inserted state with η_{min}=10.7Å while waiting for the CPD and partner bases to eject from the DNA duplex during the Rad4-DNA interaction. The free energy profile for the β-hairpin insertion is changed after they are flipped out, shifting the energy minimum to $\eta_{min} \sim$ 3Å and enabling simple insertion of the β-hairpin into the lesion location. We compared the mean potential energies computed from the unbiased MD runs of these two energy-minimum states to assess the relative stability of these two states. Indicating greater stability of the inserted state with flipped out partner bases, the computed mean potential energy of the deinserted state was 0.054 kcal/mol higher than that for the β-hairpin inserted state. In one case, the β-hairpin insertion is an uphill (ascend) process on the free energy profile with a free energy difference ΔF_I = $F_I(\eta = 3.1\text{Å}) - F_I(\eta_m in = 10.7\text{Å})$ of 46.77 kcal/mol, while for the other case, it is downhill (descend) process with ΔF = $F_E(\eta = 10.7\text{Å}) - F_E(\eta_m in = 3.1\text{Å})$ of 41.81 kcal/mol. It is tempting to interpret ΔF_E = 41.81 kcal/mol and ΔF_I = 46.77 kcal/mol as deinsertion and insertion energies, respectively, of the β-hairpin from or into the DNA duplex. However, caution must be exercised while interpreting ΔF_E and ΔF_I for the following reason. The actual deinsertion energy is the free energy difference between the inserted state and the intra-helical deinserted state, but $F_E(\eta = 10.7\text{Å})$ corresponds to the extra-helical deinserted state. A more meaningful quantification of the deinsertion energy is the absolute difference between these energies, which is equal to 4.96 kcal/mol.

3.3.2 CPD partner base flipping

We investigate the following two successive base flipping routes. The BHD3 β-hairpin inserted in the lesion site of the of the Rad4-DNA bound crystal structure, leaving the partner bases and the CPD lesion in their respective extra-helical forms. Yet, this static structure seems insufficient to reveal whether the flipping of the partner bases and CPD precedes or succeeds the β-hairpin insertion. Examining the kinetics of base flipping both before and after the insertion of the BHD3 β-hairpin into the DNA lesion site is crucial to comprehend the process of base flipping. The problem is that the BHD3 β-hairpin functions as a significant barrier to the partner bases' entry into the intra-helical state in the post-inserted state, when it is deeply inserted into the damage site. As a result, the partner bases' and CPD's extra-helical to intra-helical transition is a rare occurrence whose timelines are outside the scope of standard molecular dynamics simulations. The partner bases and CPD can be compelled to their respective intra-helical states via biassed MD simulations. These biassed MD simulations sample high-energy configurations with biassed potentials and look at how the BHD3 β-hairpin reacts to partner base flipping and CPD. Nonetheless, it is a costly computing exercise. The flipping of the partner bases and CPD, on the other hand, can be somewhat simpler than in the post-inserted state in the pre-inserted metastable state where the BHD3 β-hairpin is about to be inserted into the damage site. Therefore, it is not possible to obtain the crystal structure of this pre-inserted form. Modeling is required for this pre-inserted state. The partner bases flipping in concert or sequentially is not obvious, which presents another difficulty (the flipping of one base is followed by the other). Even with consecutive flipping, it is unclear whether 3'-dA or 5'-dA flips first. To get to any significant conclusions on the mechanism of base flipping in the Rad4-DNA complex, we must take into account all of these possibilities. We must look into the coordinated and sequential flipping of partner bases in both the post-inserted and pre-inserted phases in light of all these potential outcomes. We have not examined concerted flipping in this study since it is highly improbable and lacks experimental support. To get to any significant conclusions on the mechanism of base flipping in the Rad4-DNA complex, we must take into account all of these possibilities. We must look into the coordinated and sequential flipping of partner bases in both the post-inserted and pre-inserted phases in light of all these potential outcomes. We have not examined concerted flipping in this study since it is highly improbable and lacks experimental support. Four umbrella sampling studies were carried out to ascertain the order of the partner bases 3'-dA and 5'-dA flipping events: (a) 3'-dA flipping immediately following de-insertion (Model B) to determine its fully intra-helical conformation, denoted by the collective variable δ ; (b) 5'-dA flipping on the de-inserted state having an intra-helical 3'-dA with collective variable (Model C) γ ; (c) 5'-dA flipping immediately following de-insertion (Model B) to determine its fully intra-helical conformation represented by CV γ' ; and (d) 3'-dA flipping on the de-inserted and 5'dA flipped in model (Model D) denoted by CV δ'

3.3.2.1 3'-dA Flipping Before 5'-dA Flipping in Deinserted State (Model B)

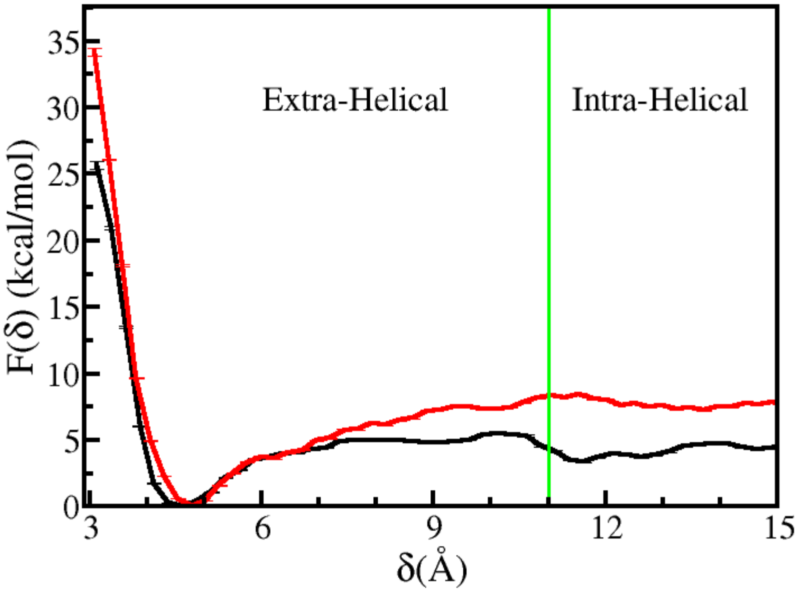


Figure 3.6: PMF profile associated with the flipping of 3'-dA is shown as a function of δ for Model B (black) and Model D (red).

The calculated free energy profile as a function of δ is shown for the Model B in Figure 3.6 (black). A global energy minimum is observed at $\delta_{\min}$ =4.57 Å, which corresponds to the extra-helical state of 3'-dA in which 3'-dA aromatically stacks with PHE434 residue of Rad4. For $\delta < \delta_{\min}$, the free energy steeply increases due to the steric clashes between 3'-dA and the key residues at the BHD2/BHD3 groove of Rad4. For $\delta > \delta_{\min}$, the interactions of 3'-dA with Rad4 gradually weaken with increasing δ and F(δ) plateaus around 4 kcal/mol at higher δ values ($\delta >$ 11.5 Å). The observed plateau region

($11.5 < \delta < 17$) corresponds to the intra-helical state of 3'-dA, where it primarily interacts with the neighbouring nucleotides of the DNA.

3.3.2.2 5'-dA Flipping After 3'-dA Flipping in Deinserted State (Model C)

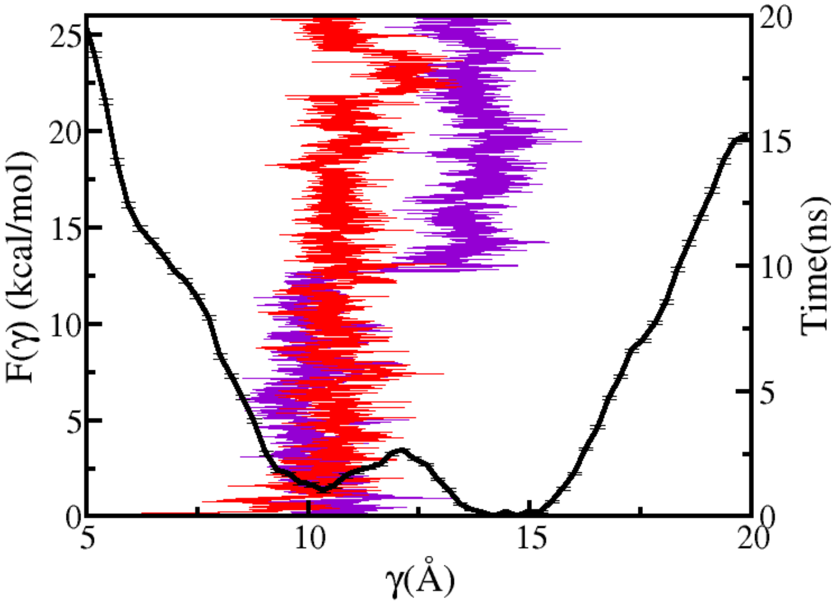


Figure 3.7: PMF profile associated with the flipping of 5'-dA is shown as a function of γ for Model C. The time series of γ obtained from two unbiased MD runs (red and violet) are also shown.

The calculated free energy profile, F(γ), associated with the flipping of 5'-dA in Model C is shown in Figure 3.7. Throughout this flipping simulation, 3'-dA remained in its intra-helical state F(γ) exhibits two energy minima; a global minimum at γ =14.2 Å and a second minimum at γ = 10.4 Å, which is 1.1 kcal/mol higher in energy than the global minimum. These two minima are separated by an energy barrier, which is located at γ = 12.1 Å, of 3.19 kcal/mol. In the most-stable global minimum conformation at γ =14.2 Å, 5'-dA is in its intra-helical state and it aromatically stacks with the neighbouring base dA. Figure 3.7 also shows the time series of γ obtained from independent 20 ns unbiased MD trajectories starting from different initial structures with different γ values. In one of these unbiased simulations, the system remained in the energy basin around the metastable minimum at γ =10.4 Å throughout the trajectory, whereas both the energy minima are sampled by the system during the course of the other simulation. As these barrier-crossing transitions between energy minima are infrequent and rare in unbiased MD simulations, these trajectories cannot be readily used to calculate statistical measures pertaining to these transitions. However, the existence of two energy minima is corroborated by the MD-derived time-series of γ . Figure 3.7, production runs of 20ns were performed on structures with different γ values and from the Figure 3.7 sharp change in γ value can be obsereved in the red curve, implying the sudden change in conormation. Also, the tendency for a conformation to stay in the metastable state is observed through the violet curve. For $\gamma > \gamma_{\mathrm{min}}$ the free energy value F(γ) is seen to increase due to the distortion of the shape of DNA around 5'-dA.

3.3.2.3 5'-dA Flipping Before 3'-dA Flipping in Deinserted State (Model B)

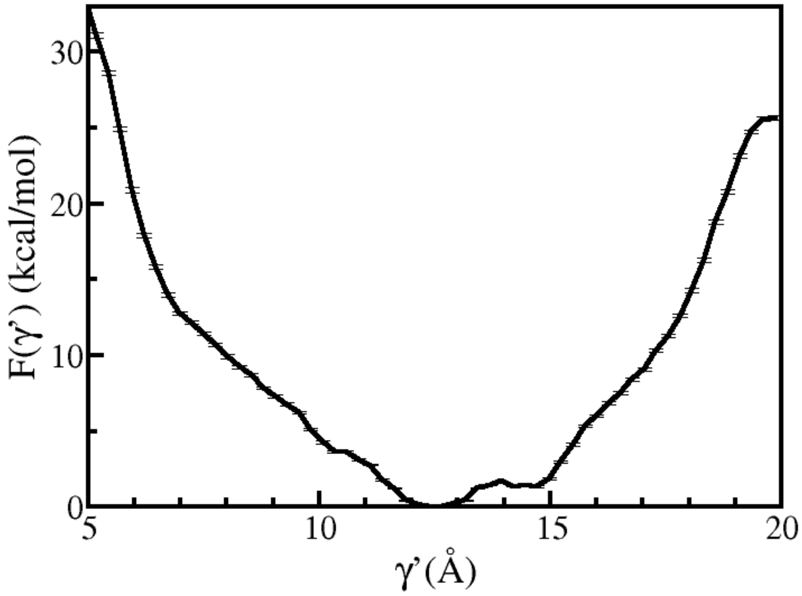


Figure 3.8: PMF profile associated with the flipping of 5'-dA is shown as a function of γ for Model B.

The free energy profile calculated using Umbrella Sampling along γ' for Model B is shown in Figure 3.8. A global minima is detected at 12.57 Å as for this value 5'-dA is in interaction with its neighbouring base A18$_{\mathrm{u}}$. This point will be hereon referred to as γ'_{min} . The γ' value plateaus in the region 11 Å - 15Å due to the interaction of 5'-dA with its neighbouring base A18$_{\mathrm{u}}$. For $\gamma' > \gamma'_{\mathrm{min}}$ an increase in the F(γ') is recorded.

3.3.2.4 3'-dA Flipping After 5'-dA Flipping in Deinserted State (Model D)

In the Figure 3.6 (red), which displays the base flipping free energy profile for Model D, global minima is perceived at 4.76 Å as for this value 3'-dA is in aromatic stacking with PHE434. For $\delta < \delta_{\mathrm{min}}$ a steep increase in the $\mathrm{F}(\delta)$ is spotted due to the steric clashes between 3'-dA and PHE434. For

$\delta_{min} < \delta < 6$ Å a steep increase in the $F(\delta)$ is observed due to diminishing interaction between 3'-dA and PHE434. After, $\delta > 6$ Å inteactions between 3'-dA and PHE434 becomes minimal. Around $\delta = 12$Å, the profile completely transformes into a plateau as in this region 3'-dA is completely flipped inside and interacts with it's neighbouring nucleotide.

3.3.2.5 Order of flipping

The most likely sequence of partner bases being extruded from the Rad4-DNA complex can be determined by comparing the free energy profiles for the flipping of the 5'-dA nucleotide after (Figure 3.7) and before (Figure 3.8) the flipping of the 3'-dA. Energy minima are seen in both of these free energy profiles at γ values corresponding to the intra-helical area ($\gamma > 9$ Å), but the characteristics of these profiles vary depending on whether the 3'-dA nucleotide is located intra- or extra-helical. It is clear from this that when the BHD3 β-hairpin is removed from the lesion site, 5'-dA prefers to be intra-helical. The observed discrepancies in the free energy profiles can be attributable to the fact that when 5'-dA flips after 3'-dA, it can stack with 3'-dA favourably in the intra-helical state. When 5'-dA transitions to a flipped out state, or $\gamma \sim 10$Å, the Δ F value, shown in Figure 3.7, is around 3 kcal/mol lower than that seen in Figure 3.8. This suggests that when 3'-dA is flipped in, 5'-dA flips out more readily. Also, it was found that 5'-dA had more freedom to translate, flip, and change its conformation from being entirely extra-helical to intra-helical during numerous unbiased and biassed simulations than 3'-dA. This finding is also in line with the energy profile shown in Figure 3.7, where the energy gap between the flipped in state (minima) and the flipped out state of 5'-dA is 1.3 kcal/mol (meta stable state around γ = 10Å). While the energy difference between these two 3'dA states in Figure 3.6 is approximately 5.1 kcal/mol. This is because 5'-dA and MET376 have relatively weaker positive interactions than 3'-dA and PHE434, which are in aromatic stacking with each other. So, when viewing this reaction from a different angle and in the direction of forward motion, it may be claimed that 5'-dA flips first and 3'-dA subsequently. These findings support earlier studies that claim it is energetically preferable for 5'-dA flipping to occur before 3'-dA flipping [86, 189, 190].

3.3.2.6 Order of BHD3 β-hairpin Insertion and CPD partner base flipping

We computed the free energy profiles for β-hairpin insertion for both the extra-helical and intra-helical conformations of the partner bases as well as the CPD lesion in order to investigate the relationship between BHD3 β-hairpin insertion and base flipping of partner bases. In Model A, where both the partner bases and CPD are expelled from the DNA duplex, we initially computed the free energy profile for the insertion of the BHD3 β-hairpin. Then, we estimated F(η) for a Rad4-DNA model (Model F) in which the partner bases and CPD are both forcibly introduced into the DNA duplex. In this case, the latter model assumes that the β-hairpin insertion comes before the flipping of partner bases and CPD and the former model assumes the reverse. The computed F(η) from these two models is compared in Figure 3.5. For the model when the partner bases and CPD were in their extra-helical conformations, F(η) reveals an energy minimum at η = 3.1 Å (later, this value of ins will be referred to as η_{min}). The energy minimum is located closer to the value of ins calculated from the Rad4-DNA complex experimental crystal structure (η_{xl}). In contrast, the energy minimum on F(ins) for the alternate model, in which the partner bases and CPD are intra-helical, is situated at η_{min}=10.7Å, 5.8Å away from the value of the crystal structure. Also, compared to the earlier model, the latter model's energy basin is larger and more lopsided. These findings imply that in the latter model compared to the former model, the BHD3 β-hairpin is relatively more dynamic close to the energy-minimum configuration. The minima for black curve corresponds to an inserted state, whereas red curve rests at a higher value, corresponding to a de-inserted state. This implies that the structure wants to be in a de-inserted state when CPD and its bases are fully flipped inside, while it prefers to be in an inserted state when these bases are fully flipped outside. These findings suggest that during Rad4's lesion detection process, the CPD and its companion bases must be flipped out to allow for the insertion of the BHD3 β-hairpin.

3.3.2.7 Flipping of Partner Bases in Inserted State

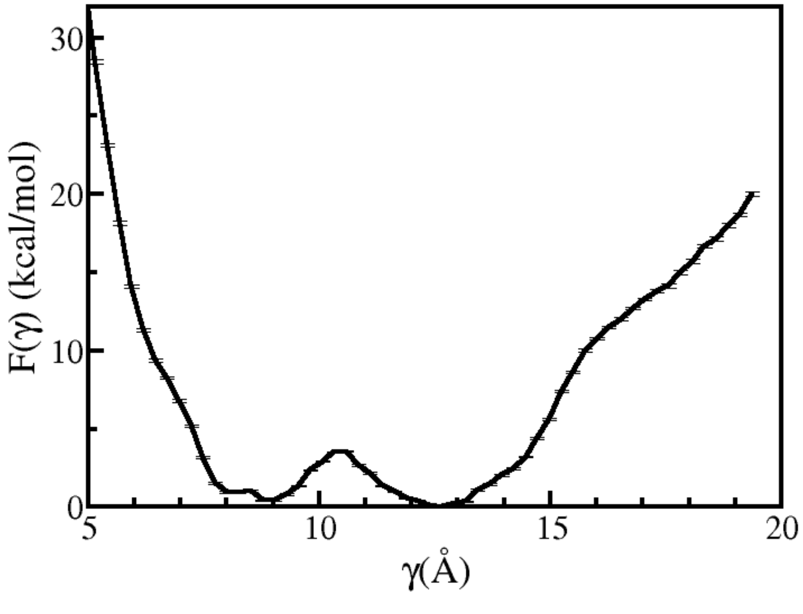


Figure 3.9: PMF profile associated with the flipping of 5'-dA is shown as a function of γ for Model A.

To further confirm these results mentioned in previous section, flipping study of 3'-dA(δ , Figure 3.10) and 5'-dA(γ , Figure 3.9) was performed on Model A, with BHD3-β hairpin inserted.

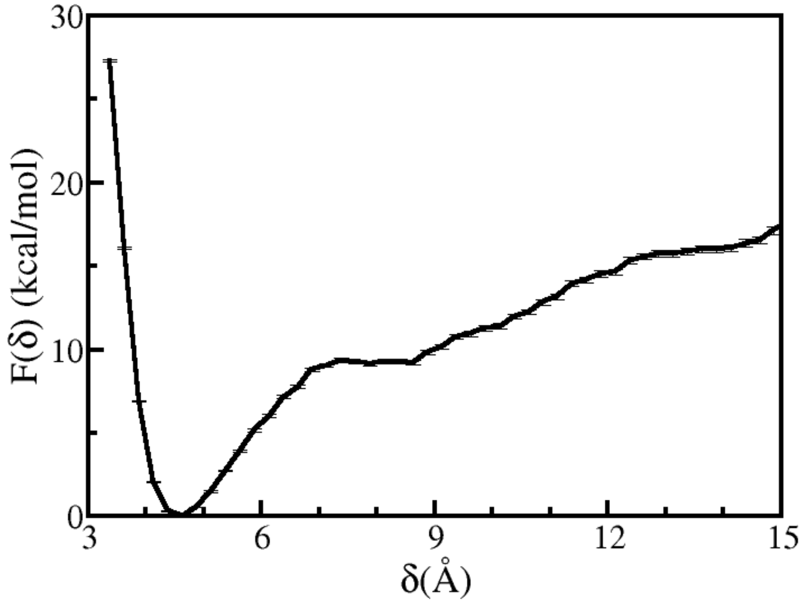


Figure 3.10: PMF profile associated with the flipping of 3'-dA is shown as a function of δ for Model A.

Comparing Figure 3.10, Figure 3.6 (black) and Figure 3.6 (red) shows that for the current inserted model, the free energy rises with increasing δ for δ larger than 9 Å, whereas for the deinserted version, the free energy stagnated in this range of δ. The present model's increased energy for δ higher than 9 Å can be attributed to steric collisions between the residues in the BHD3 β-hairpin and 3'-dA. Two iso-energetic minima are visible in the free energy profile for the flipping of 5'-dA, one at γ = 8.85 Å and the other at γ = 12.58 Å, and they are separated by a barrier of 3.2 kcal/mol. The minimum at 8.85 Å is 1.12 Å distant from the experimental value, given that the value of γ derived from the experimental crystal structure is 7.73 Å. It would seem that neither the extra-helical state nor the intra-helical state of 5'-dA correspond to these minima. The BHD3 β-hairpin at the lesion site prevents the intra-helical state from being attained. The 5'-dA base appears to be more dynamically flexible than the 3'-dA base based on the shallow energy basins surrounding these minima.

In conclusion, the results show that the flipping out process happens sequentially, with 5'-dA flipping out before 3'-dA, and that the flipping out of CPD and its partner bases is required to allow for BHD3 β-hairpin insertion during lesion detection by Rad4.

3.4 Conclusion

Many harmful effects, such as cancer and several genetic abnormalities, can result from DNA damage. The DNA repair proteins protect genomic integrity by identifying and accurately repairing DNA damage. Skin cancer and other skin-related disorders are caused by the cyclobutane pyrimidine dimer (CPD), the most common UV-induced DNA damage. A crucial repair protein called Rad4/XPC detects, investigates, and confirms the presence of CPD lesions in DNA before enlisting the aid of other repair agents to undo the harm. Researchers studying DNA damage and repair have shown a great deal of interest in the intriguing subject of how Rad4/XPC locates broken bases in the genome packaged in a busy cellular environment.

According to the current consensus on the mode of action of Rad4/XPC in DNA damage recognition, three major molecular events are suggested: (a) the association of Rad4/XPC with the mismatched DNA, (b) the flipping of a pair of nucleotide bases at the damage site, and (c) the insertion of a lesion-sensing BHD3 β-hairpin into the damage site. These molecular processes turn out to be inherently slow, making it particularly challenging to analyse them using traditional molecular dynamics simulation due to the challenging underlying potential energy landscape of this complicated system.

In this study, we investigated the molecular mechanism and energetics of both partner base flipping and BHD3-β hairpin insertion processes using molecular dynamics simulations and enhanced sampling technique. Our goal was to determine whether there was any relationship or connection between these two critical steps in the DNA damage recognition and repair process. The experimental crystal structure of the Rad4-DNA complex served as the starting point for our research. We then developed seven different intermediate models of this complex, taking into account variations in the positioning of the CPD and partner bases (intra- or extra-helical) and the presence or absence of the BHD3 β-hairpin at the lesion site. We calculated the free energy profiles for these intermediary processes using the relevant collective variables since each of these models reflects a potential endpoint of an intermediate process within the overall NER mechanism. Our results provide important new understanding of the sequence of critical NER events, including partner base flipping and insertion of the BHD3 β hairpin into the DNA duplex.

The findings mentioned in this chapter indicate that it is likely that during Rad4's lesion detection phase, partner bases flip before the β hairpin is inserted. In other words, only after both partner bases are flipped outside the DNA duplex can the β-hairpin enter the lesion region. The β-hairpin could reach a metastable configuration 10.7 Å distant from the lesion site when the partner bases were compelled to stay intra-helical during hairpin insertion. It was discovered that the flipping of the partner bases occurred sequentially, with 5'-dA extruding before 3'-dA. Furthermore, it was discovered that 5'-dA had higher conformational flexibility than 3'-dA, with the former showing more conformational diversity. The aromatic stacking interaction between 3’-dA and Rad4’s PHE434 is what causes the reduced structural diversity of 3’-dA in the extra-helical state.

Chapter 4

6-4PP lesion recognition by the DNA Damage Sensing Protein Rad4/XPC: Energetics and Mechanism

4.1 Introduction

In the preceding chapter, we delved into the mechanism through which Rad4 identifies CPD lesions in DNA. In the present chapter, our focus shifts to the exploration of another form of DNA damage called (6-4) photoproduct (6-4PP). CPD and 6-4PP are both common types of DNA damage induced by ultraviolet (UV) radiation. These lesions exhibit unique chemical properties. Contrary to CPD (cyclobutane pyrimidine dimer), which is a result of linkages between the carbon atoms at positions 5 and 6 of one pyrimidine base and the carbon atoms at positions 6 and 5 of the adjacent pyrimidine base, forming a cyclobutane ring structure, the formation of 6-4PP (6-4 photoproduct) involves a covalent bond between the carbon atom at position 6 of one pyrimidine base and the carbon atom at position 4 of the subsequent pyrimidine base. When compared to CPD, 6-4PP has a distinctive structure and set of characteristics due to its different chemical bonding patterns (Figure 4.1). As a consequence, their effects on DNA's structure and behaviour are expected to differ. Specifically, the extent of distortion induced in the DNA surrounding the damage site is likely to vary between these two types of lesions. Moreover, due to the presence of a single covalent bond between T-T pairs in 6-4PP, it is anticipated to exhibit greater flexibility, whereas CPD, with two covalent bonds between T-T pairs, is comparatively less flexible. Given these disparities in the chemical properties, structure, and dynamics of these lesions, investigating the variations in how a DNA damage repair protein detects them becomes an intriguing objective of this research. The primary aim of this chapter is to explore these dissimilarities in the mechanism by which a DNA damage repair protein recognizes CPD and 6-4PP lesions. Importantly, our prior examination of DNA containing CPDs has yielded valuable insights into the sequence of events involved in Rad4-mediated damage recognition. In this chapter, we will carry out similar analyses focusing on DNA that contains the 6-4PP lesion.

The 6-4PP lesion causes more significant structural distortion in DNA compared to CPD. 6-4PP containing DNA duplex exhibits 44° bend making it more susceptible to damage recognition and assisting

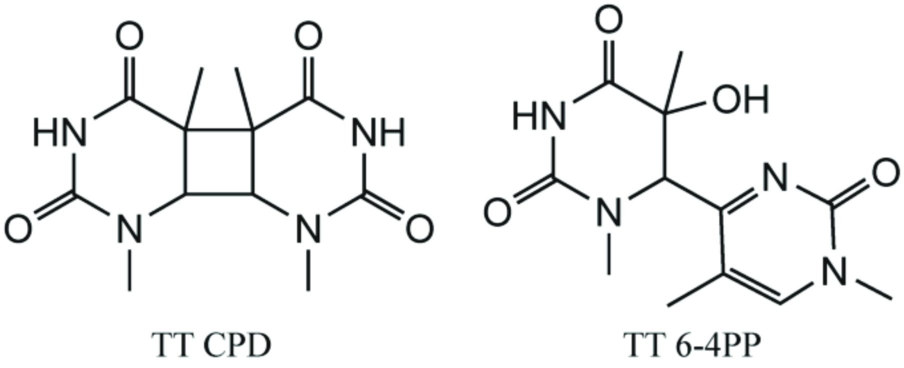


Figure 4.1: Structure of UV induced pyrimidine dimers CPD and 6-4PP adapted from [193].

in the repair bending whereas in a bound Rad4-DNA complex the bending is found to be 25° [84, 194]. Furthermore, the repair rates for the 6-4 lesion are higher compared to those for CPD. Due to these distinctions, the repair enzymes demonstrate a favourable recognition of the 6-4 lesion over CPD.

Previous studies propose that after the association of Rad4, BHD3-β hairpin inserts in the cavity that flipped-out bases leave behind.[84] This suggests that 3 key events that take place in the early damage recognition phase are the Association of Rad4 to DNA, flipping of bases and insertion of β hairpin in the damage site. This work brings light on the order and energetics of these events and make a comparative analysis with CPD repair mechanism.

4.2 Materials and Methods

4.2.1 Models

4.2.1.1 Pre-association Encounter Complex

A state of the system where Rad4 and the damaged DNA are yet to bind is referred to as the pre-association encounter complex. As an intermediate step between dissociated reactants and final bound complexes, it can be considered as a transition state. Since the crystal structure of this state wasn't available, we first constructed a model of this state using the following protocol: It was decided to construct a canonical B-DNA that followed the desired sequence (Figure 3.1) using Nucleic Acid Builder(NAB). Major parts of this sequence were taken from a damaged DNA crystal structure Figure 3.1) and the rest were added (Orange coloured nucleotides in Figure 3.1). Our modelled DNA contains two consecutive thymines replaced by a 6-4PP lesion derived from a crystal structure of an unbound 6-4PP-containing DNA (PDB ID: 1CFL).[68] Furthermore, the nucleotides coming at 19 and 20 on the undamaged strand (Nucleotides in green colour in Figure 3.1) will be referred to as 5'-dA and 3'-dA, respectively.

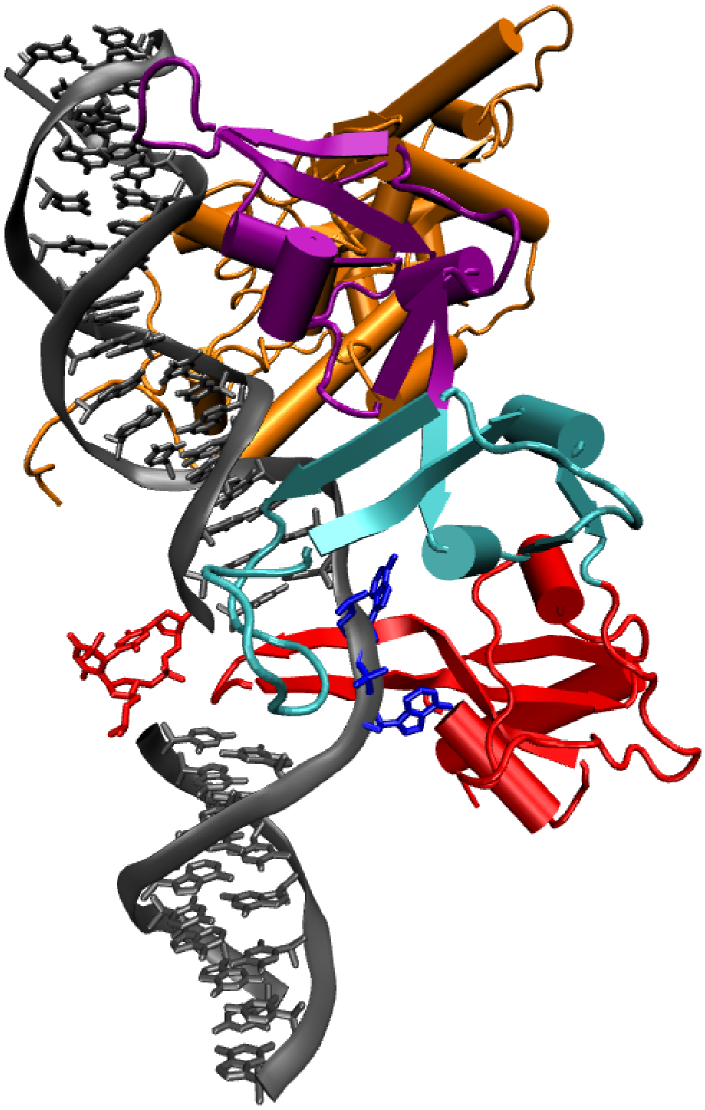


Figure 4.2: Model of 6-4PP-containing DNA-Rad4 post-recognition complex (PRC). Color code: TGD (orange), BHD1 (purple), BHD2 (cyan), BHD3 (red) domains of RAD4 and the 6-4PP (red) and its partner adenine bases (blue) of DNA (grey). The image was generated using VMD [178].

Rad4 was modelled using the crystal structure of DNA-free apo-Rad4 (PDB ID: 2QSF). In order to model the missing residues in apo-Rad4, we used the protein-Modeller program [86, 179]. We then proceeded to bind the apo-Rad4 and unbound lesioned DNA appropriately to construct the pre-association encounter complex. In the first stage of the docking, Rad4-bound DNA from the open complex was lined up with the DNA harbouring the unbound lesion (PDB ID: 6CFI). Second, the DNA-bound Rad4 of the open complex was aligned with both the TGD and BHD1 of the apo-Rad4. In the current study, this state where both apo-Rad4 and the unbound lesioned DNA are aligned to provide maximum alignment with the final open complex is used as a model of the pre-association encounter complex.

4.2.1.2 Associated Open Complex

The final related open complex was modelled after the crystal structure of the Rad4-DNA complex (PDB ID: 6CFI). The β-hairpins of Rad4's BHD2 and BHD3 are placed into the minor and major grooves of the DNA close to the lesion in this bound condition, and the subsequent adenine partner bases on the undamaged strand are fully expelled from the DNA duplex. To replicate the DNA sequence in the pre-association encounter complex, the DNA sequence extracted from this crystal structure was expanded to a 28-base pair sequence. The 6-4PP lesion was added to the open complex using the same procedure that was previously utilised to construct the pre-association encounter complex since the coordinates of the 6-4PP lesion were not confirmed in the crystal structure of the open complex. Furthermore, using the Swapna module of UCSF Chimera, each of the two mismatched thymine partner bases opposite the lesion in the crystal structure of the open complex was changed into adenine bases. [180, 181]. Finally, a sequence similar to the Figure 3.1 was obtained with the only change being T-T represents 6-4PP lesion instead of CPD. The main goal of selecting this precisely matched lesion-containing DNA is to explore the sole roles played by the lesion in this procedure and to eliminate the effect of mismatch in Rad4-DNA binding. Figure 4.2 represents the final structure obtained.

4.2.1.3 Metastable state structures

Some metastable state structures were built as a checkpoint in this study. These checkpoints mark the end of a phase in the reaction of Rad4 detecting the lesion site. These checkpoints are used to perform US for the next phase. The metastable state structures selected for this study are models B, C, D, E and F in Figure 3.4. **Model B**, a metastable state representing de-inserted conformation was formed using the algorithm mentioned in the **Models of the Metastable State** Section on de-insertion US. **Model C** was chosen such that the partner bases and 6-4PP are flipped inside representing a state just after the association of RAD4 to the lesioned DNA. **Model D and E** were formed such that 5'dA and 3'dA were flipped outside, respectively. These structures represent the state of the DNA after one of the partner bases have flipped outside. Both of these structures were formed from the US study of 3’dA and 4’dA flipping, respectively.

The models (Post Recognition Complex and Association complex) created for this chapter follow a similar procedure to that of Chapter 3 with the only difference being that the CPD lesion is replaced with 6-4PP.

The sequence used was the same as that in the 2019 breakthrough study of the 6-4PP containing DNA [84]. The lesion site as reported in this sequence was used for the constructed DNA as well. Hence, the two consecutive thymines located on the damaged strand were replaced by a corresponding lesion structure. The structures for the lesions CPD and 6-4PP were obtained from PDB ID: 1T4I and PDB ID: 1CFL, their respective crystal structures [67, 68].

The 1T4I structure was obtained using X-ray diffraction at a resolution of 2.5 Å and the 1CFL structure from the usage of solution NMR.

Apo-Rad4 was obtained from the crystal structure of apo-Rad4(PDB ID: 2QSF), followed by the modelling of missing residues using the Modeller [86, 179]. It was finally docked onto the DNA by positioning its TGD and BHD1 domains as per their conformation in the crystal structure of the ’open’ complex. These open complexes were the CPD containing mismatched DNA (PDB ID: 2QSG) and the 6-4PP containing mismatched DNA(PDB ID: 6CFI) for the lesions CPD and 6-4PP respectively [86, 189]. As for the PEC, the open complexes were taken as the reference structure.

4.2.2 Molecular dynamics simulation

AMBER 2018.10 simulation package was used to simulate the all-atom molecular dynamics simulations of the models with ff14SB force fields for the protein, ParmBSC1 force fields for DNA, and TIP3P model for water molecules. [107, 182–185] The partial charges of atoms in the 6-4PP lesion were assigned using the Antechamber module of AmberTools19 and the remaining force field parameters of the lesion were adopted from the general Amber force field (GAFF) [186]. The SHAKE algorithm was used to constrain the lengths of all hydrogen-atom bonds. [129] There was a 20Å of water padding around each of the DNA-Rad4 model complexes. The boundary conditions were three-dimensional and periodic. A distance cutoff of 10 Å was used for the van der Waals interactions. Electrostatic interactions for long ranges were handled by the particle mesh Ewald (PME) [128] approach with a tolerance of 0.00001, direct space cut-off taken at 10 Å, the Ewald coefficient at 0.27511 and the interpolation order fixed at 4.

A strong harmonic constraint was applied to the crystallographically resolved atoms of the DNA-Rad4 complex to hold them near their experimentally resolved positions during the initial phase of energy minimization in order to maintain the overall structural integrity of the complex. In the complex, unresolved atoms whose coordinates were guessed were subjected to a weak harmonic constraint. A subsequent stage of energy minimization involved removing the harmonic constraints from the unresolved atoms, while retaining those from the resolved atoms. All stages of energy minimization did not constrain the water molecules or counter ions. Following equilibration for 0.02 ns in the NVT ensemble at 300 K, the energy-minimized configurations were simulated at 300 K for 2 ns in the NPT ensemble at 1 bar, while retaining harmonic constraints on the resolved atoms. After all constraints were removed, whole systems were energy minimized for 0.02 ns and then equilibrated for 2 ns in the NVT ensemble and NPT ensemble respectively. There were 20000 steepest descent steps and then 20000 conjugate gradient steps with a convergence tolerance of 10^{-4} kcal mol^{-1} $\mathrm{Å}^{-1}$ for all minimization runs. Integrating the equations of motion took place with a time step of 2 fs using the velocity Verelet algorithm.[100] A Berendsen barostat was used to maintain the pressure at 1 bar, while a Langevin thermostat with a collision frequency of 1 ps^{-1} maintained the temperature at 300 K.[115, 187]

4.2.2.1 Umbrella Sampling

Since key molecular dynamics of NER are likely to occur at longer timescales than available timescales in conventional MD simulations, the umbrella sampling method was used for quantifying the energetics associated with conformational changes in Rad4-DNA complexes during NER that were captured. As we describe in the further sections, three primary variables were used in the present study: Rad4-DNA association, flipping of partners' bases and β-hairpin insertion.

4.2.3 Collective variable for β-hairpin insertion:

In order to identify the insertion of the β-hairpin of the BHD3 domain of Rad4, we employed a distance-based CV, η. The distance between the sugar rings of the neighbouring bases of 64PP and its partners' ($\mathrm{A18_u}$, $\mathrm{G21_u}$, $\mathrm{T18_d}$, $\mathrm{C21_d}$) centre of mass (COM) and COM of backbone heavy atoms of all the residues in BHD3-β hairpin's is η (Figure 3.3a). Equilibration and production runs were calibrated with 75 kcal/mol$\mathrm{Å}^{-2}$ and 5 kcal/mol$\mathrm{Å}^{-2}$ harmonic force constants. For umbrella sampling, η was varied by 0.5 Å steps between 1Å and 22Å. 4.947Å is the experimental value of η . A cavity is created by the flipping out of 3'-dA and 5'-dA which facilitates the insertion of BHD3-β hairpin.[189] However, the CV was based solely on distance without any restriction on the direction of BHD3-β hairpin movement.

4.2.4 Collective variable for base flipping

We defined a distance-based CV separately for each partner base that partners with the 6-4PP lesion so that we could grasp the flipping dynamics of such consecutive adenine bases.

Observations from the crystal structure of Rad4-DNA show that both partner bases (3'-dA and 5'-dA) are flipped out from DNA duplex and bound to binding pocket at the interface of BHD2 and BHD3 domains of Rad4. During this fliped-out state (or extra-helical state), 3'-dA binds favourably to some key Rad4 binding pocket residues (TYR375, MET376, and ASN377). When 3'-dA was aromatically stacked with its neighbouring G21$_{\mathrm{u}}$ base before it was expelled from the DNA duplex, these interactions did not exist. Consequently, it was considered that the distance between 3'-dA's centre of mass and the heavy atoms of the binding pocket residues would be a suitable CV to describe the flipping dynamics of 3'-dA. Henceforth, this CV will be referred to as δ as shown in Figure 3.3c. As seen in Figure 3.3c, this CV shall hereafter be referred to as δ . To equilibrate and to do production runs on δ , biasing harmonic force constants were set to 100 kcal/molÅ^{-2} and 10 kcal/molÅ^{-2}, respectively . For these runs, its value was varied from 3 Å to 16 Å with a step of 0.5 Å.

In the extra-helical state of Rad4, 5'-dA aromatically stacks with PHE434 of the BHD3 domain. Contrarily, in the intra-helical state, where it stacks with A18$_{\mathrm{u}}$, this stacking interaction is absent. To describe the flipping dynamics of 5'-dA in this study, the distance between COMs of 5'-dA and PHE434 was used as a suitable CV. As seen in Figure 3.3b, this CV will now be known as "γ ". To equilibrate and to do production runs on γ , biasing harmonic force constants were set to 100 kcal/molÅ^{-2} and 10 kcal/molÅ^{-2}, respectively. For these runs, its value varied from 3.5 Å to 20 Å with a step of 0.5 Å.

4.2.5 Collective variable for Rad4-DNA association

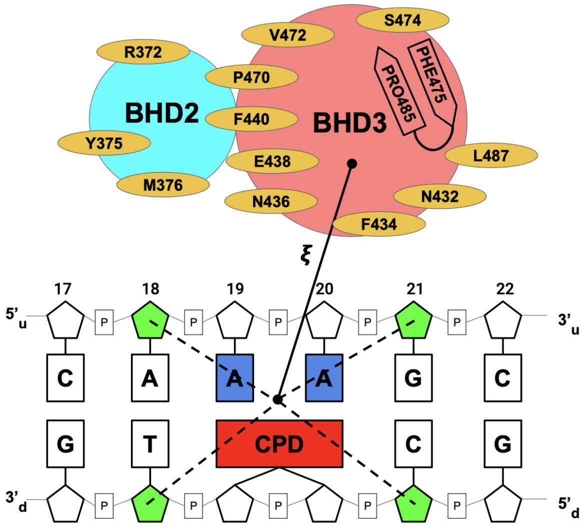


Figure 4.3: **Schematic representation of the Collective Variables used to simulate Rad4-DNA association during NER.** ξ is the distance between the COM of heavy atoms of amino acids (yellow ellipses) and the backbone heavy atoms of BHD3-β-hairpin amino acids (PHE475-PRO485) (coral), and the center of mass of the sugar rings of the neighbouring bases of CPD and their partners (A18$_{\mathrm{u}}$, G21$_{\mathrm{u}}$, T18$_{\mathrm{d}}$, C21$_{\mathrm{d}}$) (green)

ξ represents a measure of the distance that quantifies how far the center of mass of the sugar rings of the four neighboring nucleotides and partner bases (3'-dA and 5'-dA), marked in blue colour in

Figure 4.3 are from the COM of the complete BHD3 β-hairpin and the center of mass of the backbone heavy atoms in the binding pocket residues as shown in Figure 4.3, is a measure of the association between Rad4's BHD2 and BHD3 domains with the damaged DNA. It was chosen as the CV for the same due to these conditions. The biassing harmonic force constants for the equilibration and production runs were set to 100 kcal/molÅ^{-2}and 10 kcal/molÅ^{-2}, respectively. ξ was modulated from 10 to 30 Å in steps of 0.5 Å.

Each of the above events was simulated separately using umbrella sampling. In these simulations, the starting structure comes from unbiased MD simulations taken at the last frame. As part of each umbrella sampling run, the system is displaced using a harmonic biasing potential with a high spring constant (k_{eq}) so that the CV is brought to the centre of the window. To achieve this, biased NPT equilibration runs were conducted for 200 ps, for all the windows. This was then followed by a 6 ns production run in the NPT ensemble at 300 K and 1 atm pressure while subjected to a harmonic biassing potential of a weaker spring constant (k_{prod}), which is significantly smaller than k_{eq}. For various molecular events of interest, different $k_{eq}$and k_{prod}values have been chosen. The same parameters as unbiased MD runs were used in these umbrella sampling simulations, but with a supplementary restriction on the corresponding distances: (1) Employing a harmonic bias of 25 kcal/molÅ^{-2} between the COMs of bases A18$_u$ and T18$_d$, with distance being restricted at 6.06 Å. (2) A bias of 25 kcal/molÅ^{-2} between the COMs of bases G21$_u$ and C21$_d$ that are constrained at 5.85 Å. These are the neighbouring bases of the 6-4PP lesion and its partner adenines. During both the docking process and in the crystal structure, the neighbouring bases are in the intra-helicullar conformation.[189]. These neighbouring bases were restricted since the observed change in the neighbouring nucleotides is very slight and because this study has concentrated its attention to the aforementioned CVs. Additionally, because of this, the simulations used in this work are resilient to the few instances in which nearby bases adopt an extra-helical conformation.

4.2.5.1 Order of CVs

The umbrella sampling simulations were run individually for each event but progressively to look at the order of the aforementioned events and any potential coupling between them. This allows for the investigation of the impact of the bias force applied to one of the CVs on the other CVs. After an umbrella sampling is carried out independently on an event, a checkpoint/meta-stable state structure is created as mentioned in the Intermediates of Rad4-DNA Complex section of the previous chapter. Then this meta-stable structure is used for the umbrella sampling for the next event in line. Several mechanisms were investigated in order to suggest the order of CVs, and the findings are shown in the Results and Discussion section.

The following probable sequence of events has been reported in earlier investigations on NER of UV lesions [84, 188] : the DNA is first bound to the BHD2/3 domains, then the partner bases are flipped

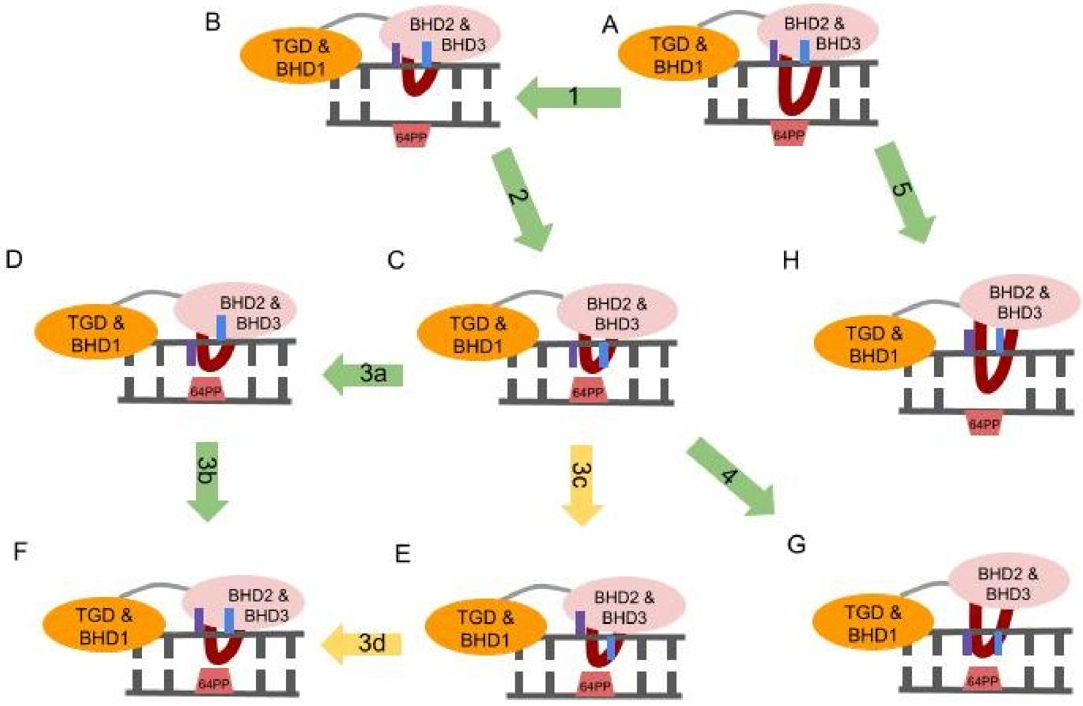


Figure 4.4: Models and sequences of events (denoted by numbered arrows) considered. (A) Rad4-DNA bound complex (B) bound complex with BHD3. (C) 6-4PP and partner bases flipped inside the active site. β-hairpin deinserted from the damage site (D, E) same as (C) except for one of the partner bases (3'-dA (blue), 5'-dA (violet)) flipped out of the DNA duplex (F) same as (C) but both partner bases are flipped out of the DNA duplex. (G) same as C but the BHD2 and BHD3 domains are dissociated from the DNA. (J) bound complex with the BHD2 and BHD3 domains are dissociated from the DNA. Transitions studied: deinsertion of BHD3 β-hairpin (1); all the bases, including 6-4PP flipped in (2); flipping out of partner bases (3a, 3b, 3c, 3d); Model A was formed by taking the most probable structure after clustering on 100ns of unbiased production run of PRC.

out, and finally the BHD3-β hairpin is inserted into the DNA. Firstly, we need to carry out a biased simulation of Rad4-DNA association, then for flipping the partner bases, and finally biased simulation for inserting the BHD3-β hairpin into DNA in order to explore the energetics of these events. In spite of that, only the crystal structure of the Rad4-DNA complex is available after recognition when the BHD2 and BHD3 domains have already been associated with the DNA. At the damaged site, the two partner bases are flipped into the extra-helical state and the BHD3-β hairpin is completely inserted into the DNA damage site. In the absence of the crystal structure of the pre-association encounter complex, modelling the sequence of events starting from the pre-association state is nondeterministic. Furthermore, the rugged energy surface of this complex system, with numerous pathways of different numbers of intermediate states separated by barriers of varying heights between the pre-associated state and the post-associated bound complex, makes determining the exact mechanism of the entire process very taxing. Thus, we reverse the entire process starting from the experimentally obtained crystal structure of the bound Rad4-DNA structure as a way of bypassing the aforementioned challenge. Hence, we are following the NER process in a backward direction in this study. Even in the reverse direction, we might encounter multiple paths. However, it is more likely that we will stay adjacent to the actual path, at least near the product stage.

With regard to the flipping dynamics of the partner bases, it is critical to determine whether the 3'-dA and 5'-dA bases flip out simultaneously or sequentially after Rad4-DNA association (the flipping of one follows the flipping of the other). There is an energetic advantage to sequential flipping of bases over concerted flipping. [86, 189, 190] There is still a lot to learn about whether flipping of 3'-dA occurs before or after 5'-dA flipping and about their associated energetics. A number of umbrella sampling simulations were conducted for different possible flipping mechanisms in order to determine the sequence of flipping of the partner bases. The flipping dynamics of both the partner bases were studied under two conditions: in one structure where the other partner base (5'-dA when 3'-dA is under study and 3'-dA when 5'dA is the one being applied bias on) was extra-helical, while in the second structure, the other partner base was intra-helical. Here, the former model essentially captures the flipping dynamics of 5'-dA or 3'-dA after the complementary partner base has flipped out, in contrast to the second model, in which 3'-dA or 5'-dA flips out before the complementary partner base. During these flipping experiments, a metastable Rad4-DNA complex with a partially deinserted BHD2 β-hairpin was used (Figure 3.4)). We will be able to better understand DNA damage recognition by Rad4 by comparing flipping energy profiles obtained from these two pathways that we have tried to model.

An umbrella sampling simulation of BHD3-β hairpin insertion into the DNA duplex of two Rad4-DNA structures was conducted in order to investigate whether BHD3-β hairpin insertion takes place before or after base flipping: in one structure, the partner bases and 64-PP were extra-helical, while in the other, they were intra-helical (Figure 3.4). A harmonic biasing potential was applied to 64-PP on the basis of its COMs distance from the partner bases COM. The results of these experiments can help us determine whether BHD3-β hairpin insertion is associated with base flipping.

4.2.6 Models of Metastable States

Multiple unbiased MD runs starting from different CV values were carried out for each of the events mentioned in previous sections in order to determine intermediate metastable states. The time series of the CV that these simulations produced were studied. A sizable portion of these trajectories were seen to cluster around a certain CV value. The RMSD-based clustering approach was used to group the structures derived from these trajectories, and the metastable state for that particular process was determined to be the core of the top-ranked cluster. The structures were realigned with regard to important nucleotides ($C17_u$ - $C22_u$ and $G17_d$ - $G22_d$) that are close to the lesion site of the damaged DNA for the RMSD calculation. The Figure 3.4 displays the several metastable states of the Rad4-DNA complex that were achieved using this method.

4.2.6.1 Force Field for 64-PP

Antechamber uses GAFF, which is compatible with the conventional AMBER force fields, to automatically calculate charges and atom kinds. The initial structure for running Gaussian from the lesion-PDB(ID: 1CFL) with a net molecular charge of **-2** is first generated using antechamber [191]. Antechamber received the corresponding log file from the optimization run to produce RESP-based partial charges.[192] The prepgen tool was used to create a mainchain file from the Antechamber output structure file, which was then used with the parmchk2 programme to check for missing force-field parameters, atoms, and atom-types.

4.3 Results and Discussion

From this point forward, we will discuss the results and inferences derived from enhanced sampling simulations based on the three major events that occur during RAD4-induced recognition. In summary, there are three events:

- BHD3 β-hairpin Insertion
- 6-4PP Partner Base flipping
- RAD4 Association

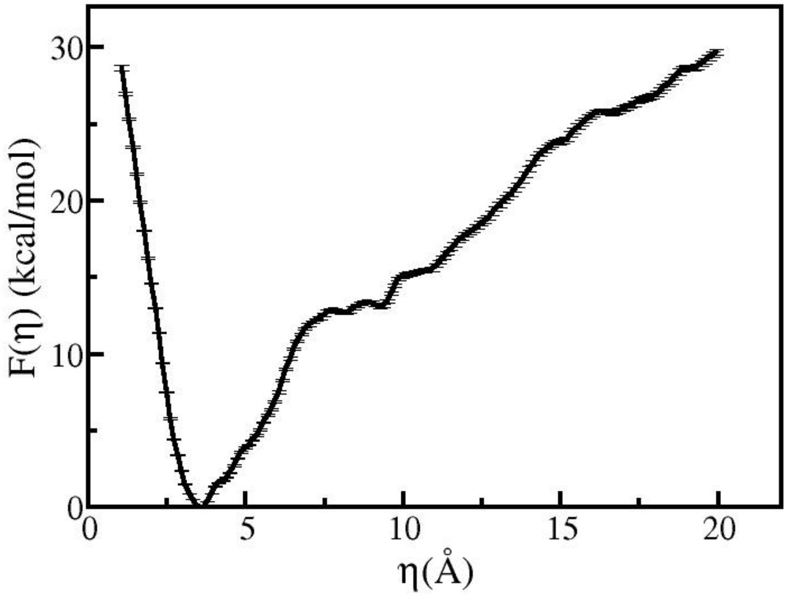


Figure 4.5: The potential of mean force for the deinsertion of the β-hairpin of BHD3 from the lesion site of the damaged DNA duplex for Model A

4.3.1 BHD3 β-hairpin De-insertion

Based on the umbrella sampling analysis, Figure 4.5 presents the free energy profile, $F_D(\eta)$, for deinsertion of the BHD3 β-hairpin from the Rad4-DNA complex. An individual minimum is observed in $F_D(\eta)$ at $\eta \sim 3.54$ Å (denoted by η_{min}), something that is relatively close to what has been measured in experimental crystal structures, $\eta = 4.94$ Å. It is also close to the value of $\eta_{min} \sim 3$ Å that was seen for the corresponding model containing CPD but the energy basin around the minima is more asymmetric for the 6-4PP model. We see a continuously increasing trend on both sides of the minima but a clear deviation from the slope that existed for 3.5 Å $< \eta <$ 7.5 Å is seen after $\eta >$ 7.5 Å. In the harmonic regime (1 Å $< \eta <$ 7.5 Å) because of its beneficial interactions with DNA, the β-hairpin seems to experience an elastic restorative force (due to hairpin-DNA van der Waals and polar contacts). This is similar to the equivalent CPD model. When dragged above this harmonic limit, the BHD3 β-hairpin appears to rupture this elastic cage at the lesion site by obstructing some beneficial interactions and displacing away from the damage site. The aforementioned slope change reflects this transition from the harmonic behaviour to the favourable interaction breaking event. In the Rad4-DNA complex, the BHD3 β-hairpin needs to be removed from the lesion site of DNA at $\eta \sim$ 7.5 Å by a minimum amount of free energy, which is ~ 12.5 kcal/mol. This energy is similar to it's CPD counterpart whose transition free energy was found ~ 13 kcal/mol. Although the transition energies are similar, the cage breaking point comes at a higher η value for 6-4PP.

A realistic deinserted meta-stable state of the system prior to the β-hairpin insertion can be found using the BHD3 β-hairpin deinsertion trajectories produced by the umbrella sampling simulations. Similar to CPD, as mentioned in the previous chapter the experimental structure of this metastable state is not accessible. To proceed with the other events in the lesion recognition, this metastable state was modelled by choosing the cluster centre of the agglomerated unbiased runs of the shoulder points. The criteria for selecting the shoulder points is the same as mentioned in the last chapter's results and discussion, but here the value of shoulder points is slightly higher. The value for shoulder points is, 9 Å < η < 12.5 Å. The ΔG value for this structure is ~ 15 kcal.mol^{-1}, which is a significantly lesser value than its CPD counterpart stated in the previous chapter and η value is equal to 10.84 Å. This model, Model B has BHD3-β hairpin de-inserted but the partner bases and 6-4PP are still flipped out.

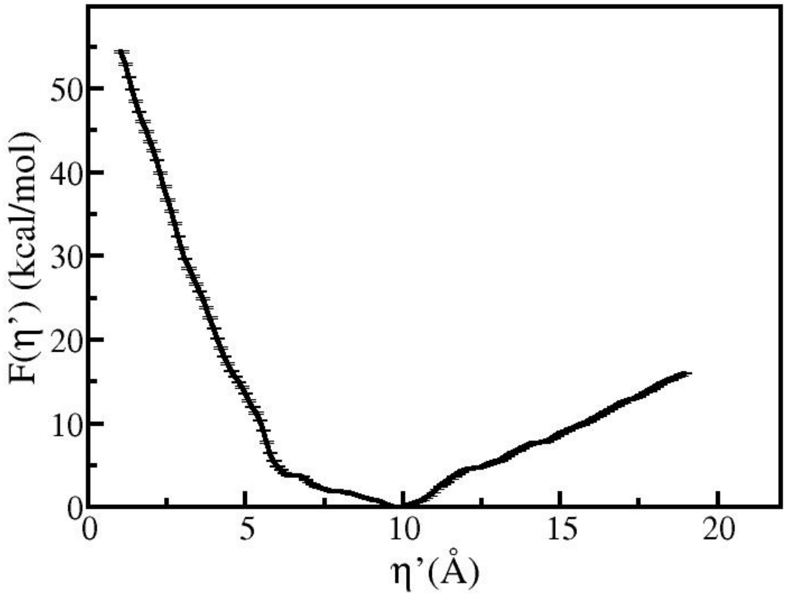
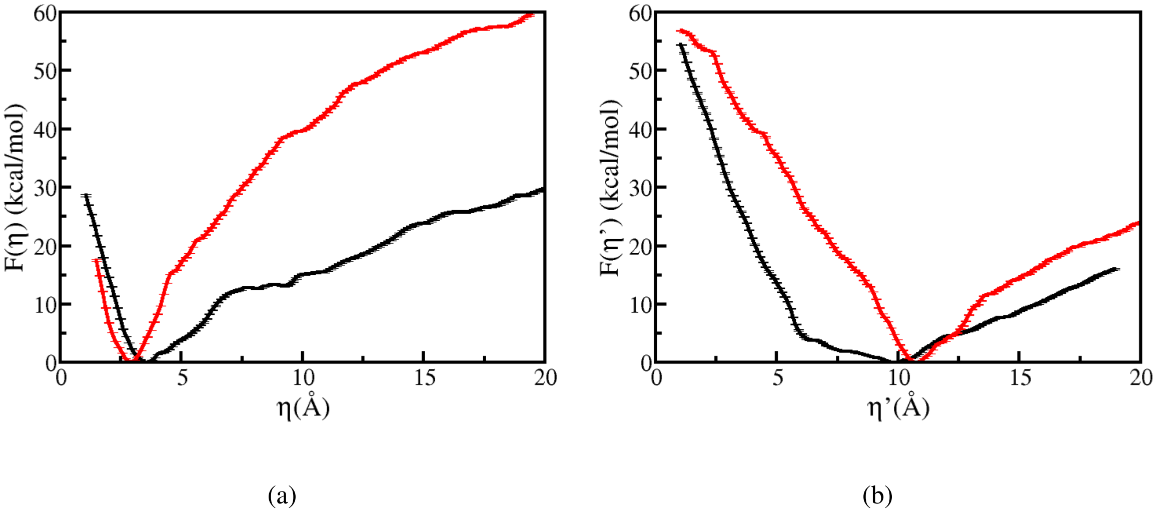


Figure 4.6: The potential of mean force for the insertion of the β-hairpin of BHD3 from the lesion site of the damaged DNA duplex for Model C

After forming this intermediate structure which represents the onset of BHD3-β hairpin insertion, both the partner bases as well as 64PPlesion were flipped into a intra-helicular state. This model, Model C, represents the state where Rad4 has associated itself with lesioned DNA BHD3-β hairpin approaching the lesion site. Umbrella sampling with collective variable η was performed on this state and Figure 4.6 was obtained as an analysis of energetics. The minima for this Umbrella sampling was observed at 9.91 Å, which is significantly higher than the crystal structure and value. Similar to the last chapter, we see that for the model containing 6-4PP lesion the model with all the bases flipped inside gives minima for a de-inserted state. We observe an increase in free energy with the decrease in η value. Here, unlike the de-insertion in open complex, the order of $F_D(\eta)$ is similar to what was observed for CPD containing model in similar conditions. This increase in free energy is due to the increasing steric clashes between the BHD3-β hairpin and lesion site where 6-4PP with partner bases are intra-helical. According to the energy difference of energy loss associated with the hairpin insertion for Model A, which is 14.9 kcal/mol compared to the energy cost required to insert BHD3-β hairpin in Model C (25.6 kcal/mol), BHD3-β hairpin prefers to remain in the de inserted state up until 6-4PP and partner bases are flipped out.



(a)

(b)

Figure 4.7: **Comparison of the deinsertion and insertion PMFs of the 6-4PP and CPD complexes.** (a) Comparison of the potential of mean force for the deinsertion of BHD3-β hairpin from the lesion site of the damaged DNA duplex for 6-4PP's Model A (black) and CPD's Model A (red). For both these models lesions and partner bases were flipped out. (b) Comparison of the potential of mean force for the deinsertion of BHD3-β hairpin from the lesion site of the damaged DNA duplex for 6-4PP's Model C (black) and CPD's Model F (red). For both these models lesions and partner bases were flipped in.

We observed the same phenomenon for both 6-4PP and CPD-containing models, further emphasising this point. It can also be interpreted that irrespective of the lesion BHD3-β hairpin is unable to insert itself in the lesion site until there is some cavity available at the active site. The shift in the energy minimum closer to zero in PMF obtained for deinsertion (Figure 4.7a) indicates that the BHD3-β hairpin demonstrates a somewhat deeper penetration into the binding pocket in the complex containing CPD compared to 6-4PP. Additionally, the energy basin surrounding the minimum is steeper for CPD complex, which is significantly shallower. These findings imply that, in contrast to the CPD complex, where the BHD3-β hairpin is more constrained, the binding pocket in the 6-4PP complex allows for relatively higher flexibility of the BHD3-β hairpin. In addition, the energy cost of removing the BHD3-β hairpin is significantly larger for CPD than for 6-4PP. But when the potential of mean force (PMF) profiles for β-hairpin insertion are compared, particularly in intra-helical complexes, it becomes clear that there are important differences between CPD and 6-4PP. The β-hairpin appears to reach its energy minimum state at distances of 10.7 Å and 10 Å from the binding pocket for CPD and 6-4PP, respectively, in circumstances where the partner bases and lesion reside within the helix. In comparison to CPD, the energy basin for 6-4PP is shallower (Figure 4.7b). Furthermore, it is clear that inserting the β-hairpin into the binding pocket is relatively easier for the 6-4PP complex than for CPD in terms of energy expenditure as suggested by the shallower basin around the minima for the 6-4PP complex in Figure 4.7b. The internal flexibility of 6-4PP may serve as a significant factor contributing to these observed differences.

4.3.2 6-4PP partner base flipping

In order to the study base flipping of partner bases, Umbrella sampling using flipping CVs, γ and δ were performed with Model B as starting point but for the same collective variables as the last chapter the partner bases were not able to flip inside. In order for this study to be consistent and as a workaround a model in which all the bases were flipped inside, Model C, was chosen. This can be considered a good starting point as this model represents the state where the lesion site must open up a cavity as mentioned in the last section. So, it can also be said that Model C is on the verge of commencing partner base flipping. By keeping the CVs same we can also gain some insight into the comparative analysis of partner base flipping in a CPD and 6-4PP lesion site, although the starting structure for them is different. Unlike the CPD study, the results obtained here are in the presence of 6-4PP, interacting with partner bases since all of them are intra-helical.

Here in order to study the mechanism of flipping of partner bases in a 6-4PP lesioned DNA, two pathways were explored using biased simulations since flipping of partner bases is on a timescale that is beyond the timescales of conventional unbiased molecular dynamics. In the first pathway, flipping of 5'-dA partner base was studied using Umbrella Sampling when 3'-dA was in an intra-helicular state. After this experiment, an intermediate state using the process mentioned in the earlier section was formed. This intermediate state had 5'-dA flipped outside interacting with the BHD2-3 pocket and 3'-dA was flipped inside. Then flipping of 3'-dA was observed using δ' on this intermediate state. In the second pathway, the study of 3'-dA flipping was done first, then after the formation of intermediate when 3'-dA was extra-helical, interacting with PHE residue in BHD2-3 pocket, and 5'-dA was intra-helical, umbrella sampling using γ' was performed to observe flipping of 5'-dA. With these pathways, it is possible to comment on the order of flipping of partner bases in a 6-4PP lesioned DNA. Similar to the previous chapter, higher γ or δ represents intra-helicular state as the partner base is far away from the BHD2-3 pocket and lower CV value represents extra-helicular state as the bases will be inside the BHD2-3 pocket away from their neighbouring bases.

As mentioned earlier four US were performed: (a) 5'-dA flipping denoted by γ on Model C, (b) flipping of 3'-dA represented by δ' on Model E, (c) 3'-dA flipping denoted by δ was studied on Model C and (d) flipping study of 5'-dA using CV γ' on Model D.

4.3.2.1 5'-dA Flipping out in Deinserted State (Model C)

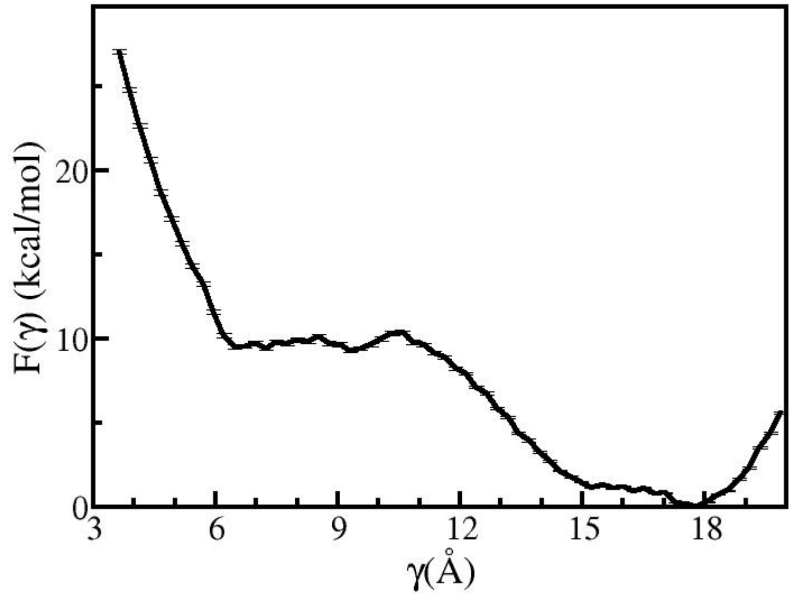


Figure 4.8: PMF profile associated with the flipping of 5'-dA is shown as a function of γ for Model C.

Figure 4.11 illustrates the free energy profile as a function of γ . A minimum is observed at γ ~17.7 Å corresponding to the intra-helical state of 5'-dA . We observe the minima for this conformation as 5'-dA in aromatic stacking with it's neighbouring bases (including 3'-dA) and positive interaction with 6-4PP. There is an increase in free energy for $\gamma >$ 17.7 Å as the more the distance between the BHD2-3 pocket and 5'-dA the more the shape of DNA distorts and favourable interactions decline. We see a constant rise in free energy from $\gamma \sim$ 17.7 Å 10.5 Å as the γ starts to break its favourable interactions with its neighbouring bases and attains an extra helical structure. There is a plateau for 6.5 Å$< \gamma <$ 10.5 Å as for these values 5'-dA is completely extra-helical and has favourable interactions with MET in the BHD2-3 pocket. But as the distance between them decreases more($\gamma <$ 10.5 Å) these interactions are replaced by steric clashes between 5'-dA and MET. Points were chosen from this plateau and a metastable state, with γ value equal to 7.89 Å was built using the algorithm mentioned in the models of metastable state section. This model was chosen as a starting point for the flipping of 3'-dA.

4.3.2.2 3'-dA Flipping After 5'-dA Flipping in Deinserted State (Model E)

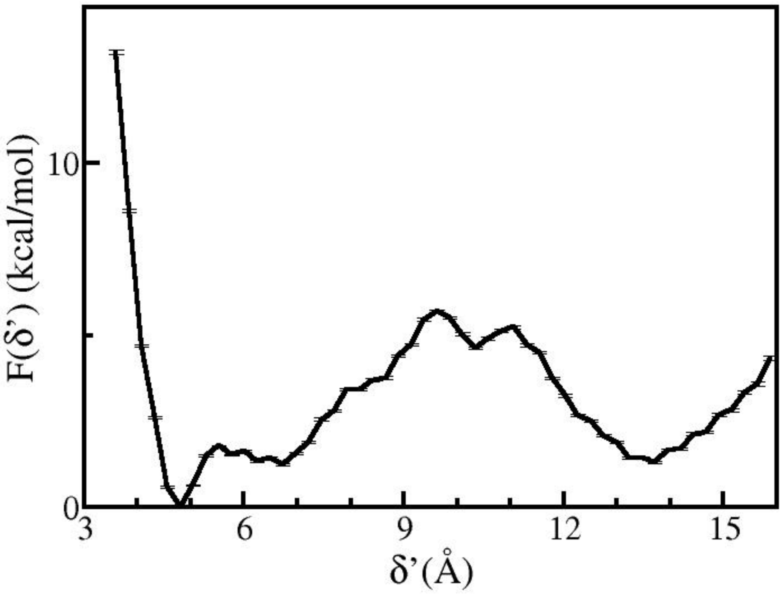


Figure 4.9: PMF profile associated with the flipping of 3'-dA is shown as a function of δ ' for Model E.

Free energy vs δ′ curve is displayed in Figure 4.9. From this figure it can be seen there are 2 minima. The global minima correspond to δ′ equal to 4.8 Å which is an intra-helical state whereas the second minima at δ′ equal to 13.6 Å correspond to an extra-helical state which is close to the δ′ = 14.8 Å value of the metastable state used. The stabilisation cause for global minima is strong aromatic stacking of 3'-dA with PHE in the BHD2-3 binding pocket. And second minima in the extra-helical state is due to 3'-dA s interaction with its neighbouring base and 6-4PP although due to 5'-dA being extra helical there is no favourable interaction there. When traversing from second minima to global minima the free energy increases and decreases again as when the 3'-dA is pulled away from the lesion site the favourable interactions with neighbouring base and 6-4PP start to, with a peak at δ′ ~ 9.6 representing breaking of all these interactions. From this point to global minima the δG value starts decreasing as it starts interacting with the BHD2-3 pocket and at the global minima it finally aromatically stacks with PHE. There is a sharp increase for δ′ < 4.8 Å as those favourable interactions transit to steric clashes as the distance between their COMs decreases.

4.3.2.3 3'-dA Flipping out in Deinserted State (Model C)

The calculated free energy for 3'-dA for Model C is displayed in Figure 4.10. The energy profile for this profile is observed for an intra-helical state at δ equal to 13.1 Å which is very close to the second

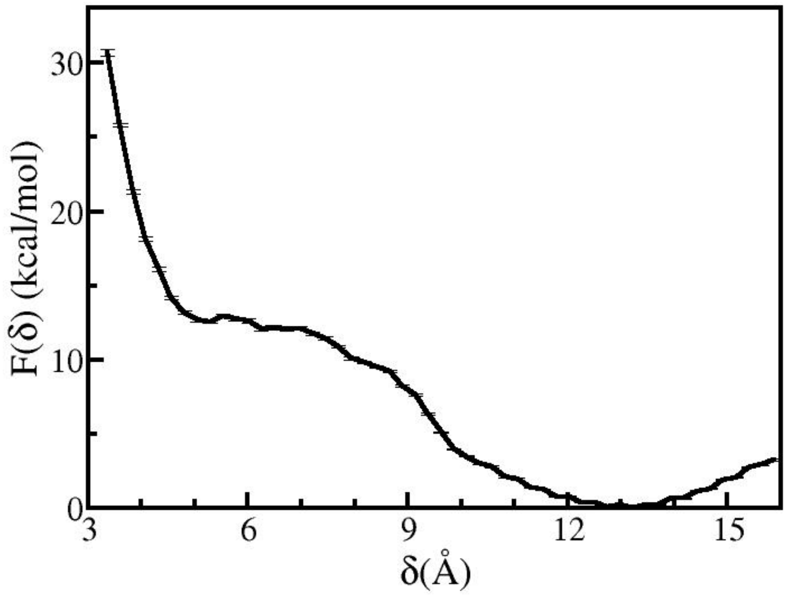


Figure 4.10: PMF profile associated with the flipping of 3'-dA is shown as a function of δ for Model C.

minima of the previous experiment. The stability of this state can be defined due to 3'-dA s positive interactions with it's neighbouring bases, including 5'-dA as it is flipped inside, and 6-4PP lesion. Then there is a constant increase in the δ G value when moving from δ ~ 13.1 Å to δ ~ 7 Å due to the breaking of the aforementioned interactions. There is a small plateau between δ equal to 7 Å and 5.1 Å due to 3'-dA coming in aromatic stacking with PHE. And similar to the previous experiment these interactions turn to steric clashes once the distance between their COMs becomes too small, 5.1 Å in this case. The plateau points we chose as a starting point for the metastable state creation algorithm. In this metastable state, which will be used for the 5'-dA flipping study, 3'-dA is flipped outside and is in aromatic stacking with PHE in BHD2-3 pocket with δ value equal to 5.48 Å.

4.3.2.4 5'-dA Flipping After 3'-dA Flipping out Deinserted State (Model D)

The calculated free energy for 5'-dA flipping for Model D Figure 4.11. This figure showcases two minima, global minima corresponding to intra-helical state with γ' ~ 17.6 Å (close to the γ' value of the metastable state, 15.42 Å) due to it's favourable interactions with neighbouring base and 6-4PP and one minimum corresponding to extra-helical state at γ' ~ 7.7 Å due to it's stabilising association with MET and BHD2-3 binding pocket. There is an energy barrier between these two states at γ' ~ 12.6 Å as at this point 5'-dA is interacting with neither it's neighbouring base nor with the BHD2-3 pocket. Similar

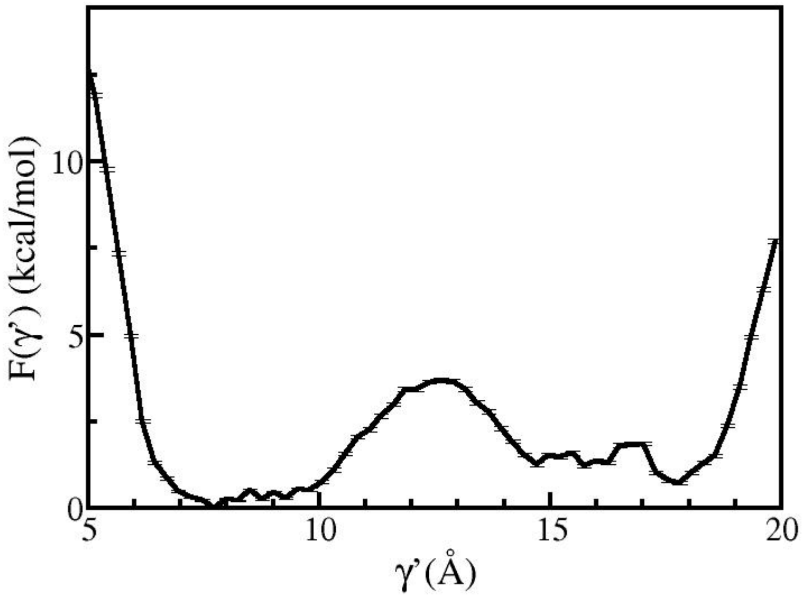


Figure 4.11: PMF profile associated with the flipping of 5'-dA is shown as a function of γ' for Model D.

to experiment(a) the free energy starts to increase when the γ' value becomes lower than 7.7 Å due to increasing steric clashes.

From experiments (b) and (d) some similarities with the previous chapter's CPD partner base flipping can be perceived, especially during the transition phase from intra to extra helical states. In both of these experiments, we observe similar minima states with similar collective variable values. For experiment (d) the energy barrier at γ' equal to 12.6, 3.66 kcal/mol is also similar to CPD's 5'-dA flipping with the value of 3.12 kcal/mol at γ equal to 12.1 Å when 3'-dA was flipped outside(similar conditions). In experiment (b) we observe minima at 4.8 Å and CPD's 3'-dA partner base flipping the minima was observed at 4.57 Å. These similarities are observed even when the direction of these pathways is opposite. In CPD's case the partner bases were flipped outside and were flipped in one by one whereas, in 6-4PP's case, the partner bases were intra-helical and flipped out one by one. The similarities despite these differences further emphasize the quality of the metastable states and pathways constructed.

4.3.2.5 Order of flipping

Comparing the free energy profiles of the four experiments in the previous section suggests the most probable order in which partner bases are extruded in the Rad4-DNA complex. For the first pathway(experiments (a) and (b)) when 5'-dA flips out first and 3'-dA later, the energy equivalent to

9.69 kcal/mol is required to reach the 5'-dA flipped out state from the completely intra-helical state and 1.24 kcal/mol worth of free energy is released by the flipping of 3'dA (although it has to cross a barrier of 4.39 kcal/mol then it proceeds to release the energy of 5.63 kcal/mol). For the second pathway(experiments (b) and (c)), where flipping of 5'-dA is preceded by extrusion of 3'-dA, it takes 12.44 kcal/mol δG to extrude 3'-dA from its most stable intra helical state and 0.65 kcal/mol is released when 5'-dA s flipping follows that of 3'-dA (here barriers energy requirement being 2.97 kcal/mol and release being 3.62 kcal/mol after crossing this barrier). In pathway 1 the total energy spent is 8.45 kcal/mol (9.69 kcal/mol for 5'-dA flipping + 4.39 kcal/mol barrier crossing energy required by 3'-dA - 5.63 kcal/mol energy released after 3'-dA crosses the energy barrier) whereas in pathway 2 the value is 11.79 kcal/mol (12.44 kcal/mol for 3'-dA flipping + 2.97 kcal/mol barrier crossing energy required by 5'-dA - 3.62 kcal/mol energy released after 5'-dA crosses the energy barrier). Therefore, pathway 1 with 5'-dA partner base flipping followed by partner base flipping of 3'-dA is more energetically favourable. It is consistent with previous studies as well as the last chapter, where the lesion was CPD, that it is more advantageous for 5'-dA to flip out before 3'-dA from an energetic standpoint. [86, 189, 190]

4.3.2.6 Order of BHD3 β-hairpin Insertion and 6-4PP partner base flipping

Similar to the last chapter, we calculated the free energy profiles for β-hairpin insertion for both intra-helical and extra-helical conformations of the partner bases and the 6-4PP lesion in order to determine whether there is a correlation between the insertion of BHD3 β-hairpin and base flipping. The first step was to compute the free energy profile for the BHD3 β-hairpin insertion in the open Rad4-DNA complex i.e. experimental crystal structure, where both 6-4PP and partner bases were removed from the duplex. For the crystal structure, we slightly deinserted the BHD3 β-hairpin to forcefully integrate both partner bases and 6-4PP into the Rad4-DNA model. BHD3 β-hairpin insertion in the former model occurs after the flipping of the partner base and the 6-4PP, whereas in the latter model, it occurs after the flipping of the base. For the former model, Figure 4.5 displays a minimum for η ∼ 3.6 Å which corresponds to an inserted state whereas in the second model, the minimum is observed at 9.91 Å, representing a de-inserted state. It is clear from these two experiments that when all the bases are in intra-helicular state the BHD3 β hairpin prefers to stay in a de-inserted state, but as we know from experimental studies and crystal structure the β hairpin insertion does take place. [189] This suggests that a cavity must be opened up at the lesion site for the BHD3 β-hairpin insertion to take place as we can clearly see in Figure 4.5 when all the bases have produced a cavity by flipping out the minimum is for an inserted state which is also close to the crystal structure value. Advocating that 644PP partner base flipping should precede the onset of BHD3-β hairpin insertion. Looking from an energetics perspective, an energy around 15 kcal/mol is released when BHD3-β hairpin inserts itself in the presence of a cavity and as mentioned in the previous section for pathway 1 total energy required(for 5'-dA flipping and 3'-dA energy barrier crossing during flipping) is 14.08 kcal/mol which is similar to aforementioned released energy. This suggests that BHD3-β hairpin and 6-4PP partner base flipping are coupled to some extent.

4.3.3 Active site and Rad4 Dissociation

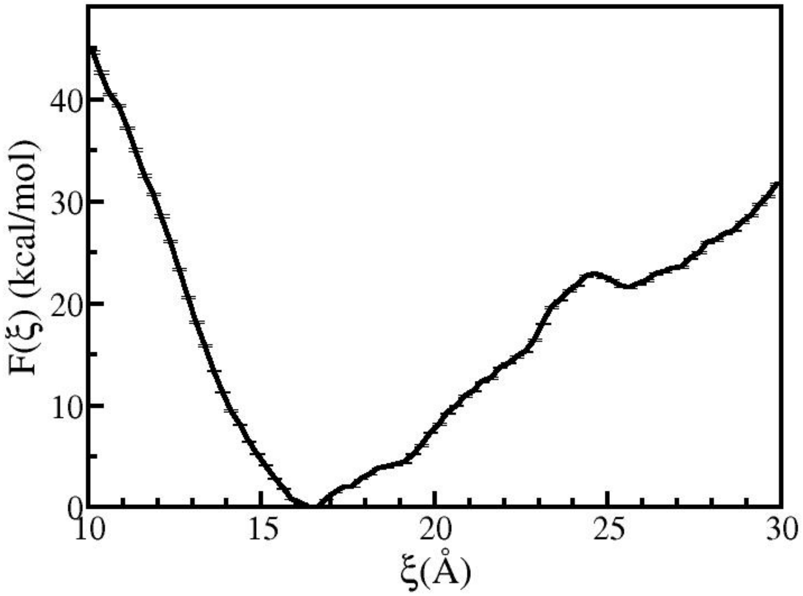


Figure 4.12: PMF profile associated with the dissociation of Rad4 from DNA shown as a function of ξ for Model C.

As a starting structure, we considered a Rad4-DNA structure with non-associated partner bases and 6-4PP, but dissociated BHD2 and BHD3 domains of Rad4. Model C fulfilled all the criteria as it represents a state just after the association of Rad4. Figure 4.12 illustrates the free energy profile, F(ξ) obtained through umbrella sampling using collective variable, ξ. In this profile, a minimum can be discerned at ξ = 16.48 present at the base of an asymmetric energy basin Å. This minimum value is far from the crystal structure value of 13.81 Å, δG value at this point being 11.8 kcal/mol. The free energy value increases for ξ < 16.48 Å due to increasing steric clashes between DNA and BHD2,3 domains. As the crystal structure value also lies in this region it can be deduced that with the insertion of flipping of partner bases and BHD3 β-hairpin insertion that Rad4 is able to associate itself more strongly with DNA. For the ξ values greater than 16.48 Å an increase in δG value is observed due to the diminishing positive interactions between DNA and BHD2,3 domains. Furthermore, a slope change at ξ = 19.2 Å can be discerned indicating that for ξ >= 19.2 Å these favourable interactions have completely been broken down.

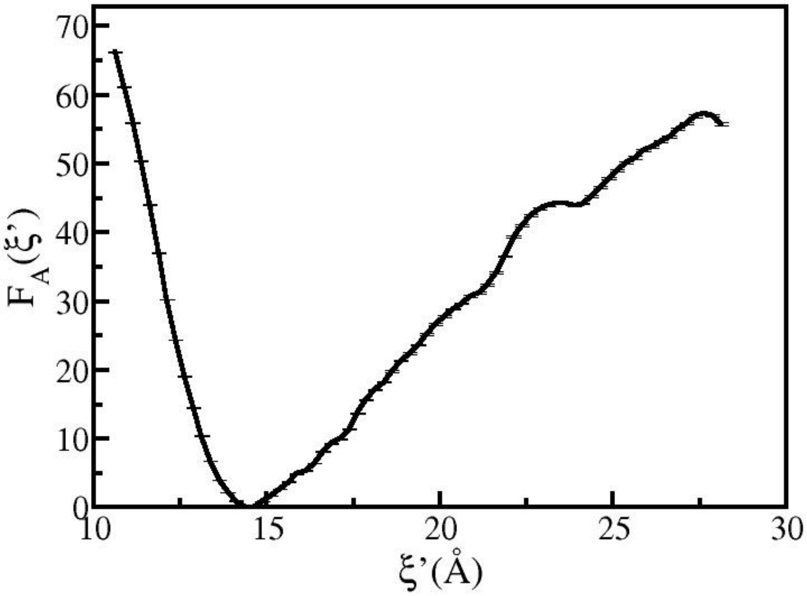


Figure 4.13: PMF profile associated with the dissociation of Rad4 from DNA shown as a function of ξ' for Model A.

Dissociation study was also performed on the crystal structure, which unlike Model C has the partner bases and 6-4PP in extra-helical state and BHD3 β hairpin inserted in the lesion site. Figure 4.13 shows the free energy for the said experiment. Here the global energy minimum at ξ = 14.48 Å which is much closer to the crystal structure value of 13.81 Å represents a completely associated state. This confirms that with the progress of the lesion recognition reaction, Rad4 binds itself more closely to the DNA compared to the initial interaction before the partner bases are flipped out.

In summary, an analysis of all computed energy profiles associated with the dissociation of Rad4 from DNA, flipping of the partner bases, and inserting the BHD3 β-hairpin in 6-4PP lesioned DNA revealed the following conclusions: first, the flipping of 5'-dA precedes that of 3'-dA even if the present lesion is 6-4PP. second, these flipping events precedes the insertion of the BHD3 β hairpin irrespective of the lesion present in the DNA.

4.4 Conclusion

As mentioned in the previous chapter, DNA damage can lead to various chronic diseases. And 6-4PP is the second most prevalent lesion found in DNA damage after CPD. 6-4PP lesion is recognised by Rad4/XPC, which then recruits other proteins to reverse the damage. The three major parts of this recognition process are (a) Rad4/XPC association with 6-4PP lesioned DNA, (b) the flipping of a 6-

4PP's partner bases and (c) the insertion of BHD3 β-hairpin of Rad4/XPC into the lesion site. Since the molecular processes in this complex system are necessarily slow, conventional molecular dynamics simulations cannot be used to investigate them. In this chapter, we computed the free energy profiles for each of these processes using appropriate collective variables and also compared them with the corresponding results obtained in the previous lesion where the lesion was CPD. Through the inspection of the energy profile of BHD3-β hairpin deinsertion, it can be concluded that significantly less amount of energy is released during hairpin's insertion in a 6-4PP lesioned DNA than a CPD containing DNA. The results in this chapter reveal irrespective of the lesion present in DNA, the 5'-dA partner base flips before the 3'-dA partner base and the flipping of bases should precede instertion of BHD3-β hairpin to open a cavity, making the insertion energetically favourable. The presence of a complex encompassing a 6-4PP lesion promotes the insertion of BHD3-β hairpin with higher ease and flexibility compared to a complex having a CPD lesion. Through the inspection of energy profiles of the dissociation study performed on crystal structure and Model C, it can be said that through the progress of later events RAD4 is able to bind itself more strongly to the lesioned DNA.

It is crucial to comprehend the structural features, repair techniques, and biological impacts of (6-4) photoproducts in order to create mitigation measures for their damaging effects on DNA integrity. Although this chapter sheds the light on recognition of 6-4PP lesion further research is needed to advance our knowledge of (6-4) photoproducts and improve our ability to protect against and repair this specific UV-induced DNA lesion.

Chapter 5

Conclusion

Many biological experimental techniques, including microscopy and biochemistry, have limitations when it comes to providing extensive information regarding the structure and dynamics of molecules at the atomic level is a significant obstacle. In addition, many biological systems, such as the interactions between proteins and DNA in the cell, are challenging or unfeasable to examine experimentally. By offering detailed data on the structure and dynamics of biological systems at the atomic and molecular level, computational approaches like molecular dynamics (MD) simulations can contribute in tackling these challenges. In-depth knowledge on processes like conformational change and ligand binding can be captured using MD approaches at a high temporal resolution, making them a perfect complement to studies of proteins and DNA that rely on their structural information. Interpreting the data and grasping the complexity of the systems, even with the assistance of these simulations, remains a significant issue. Here, enhanced sampling methods like Umbrella Sampling (US) play a huge role as it can be used to explore the activity of these molecules in regions that are unattainable under typical experimental circumstances by driving the system to energetically prohibitive regions. The fundamental idea underlying umbrella sampling is to simulate a system with a biassing potential acting along the appropriate CV, confining the system to a set of CV values. The biassing potential is generally implemented on the CV as a harmonic restraint, with the strength of the restraint selected so that the system can still explore the whole CV space while being biassed toward the region of interest. The present study focuses on finding the energetics and the mechanism behind the DNA-recognition of an unmismatched lesioned DNA using molecular dynamics and enhanced sampling methods.

The first two chapters in this thesis present the several components in a modeled Protein-DNA binding event and the computational methods employed to study their dynamics. These chapters go through the essential connections that allow DNA, proteins, and protein-DNA complexes to operate as well as their structural makeup. Furthermore, a brief overview of MD is given along with an explanation of the several computational strategies used by contemporary MD engines like AMBER.

The Third chapter explores energetics and mechanism behind the CPD lesion recognition by XPC/Rad4. The objective of this chapter is to shed light on the order and dynamics of major events, CPD partner base (including order of flipping of these partner bases) flipping and BHD3 β-hairpin insertion during recognition of CPD through energetics. The free energy profiles for flipping of partner bases and BHD3 β-hairpin in multiple plausible pathways were concieved using enhanced sampling and molecular dynamics simulations with the help of suitable Collective Variables (CV). These pathways were formed by the creation of multiple metastable/intermediate states. These metastable states were formed by choosing the cluster center of the most populated cluter formed from unbiased runs of specific points chosen from the free energy surface of a particular study. Then the metastable formed is used for next study in the pathway. For example after the BHD3 β-haripin deinsertion study, several unbiased runs were performed on shoulder points of the energy profile and then cluster center was chosen from the clustered unbiased run. Then the deinserted state obtained here was used for partner base flipping study.

An in depth inspection of the obtained free energy profiles revealed that the flipping of partner bases preceds BHD3 β-hairpin insertion during CPD recognition as some sort of cavity in the lesion site proves to be a prerequisite for insertion to move ahead. Also, during partner base flipping the pathway with 5’ dA partner flipping out from the lesion site before flipping of 3’ dA can be deemed more energetically favourable. 5’ dA partner base was found to be conformationally more flexible compared to 3’ dA as 3’ dA comes into aromatic stacking with PHE in the BHD2-3 binding pocket.

Chapter Four focuses on the energetics and mechanisms of 64PP recognition in a lesioned DNA by Rad4/XPC. The major events focussed in this chapter were association of Rad4 with the lesioned DNA, flipping of 64PP partner bases and insertion of BHD3 β-hairpin insertion in a 64PP lesioned DNA and compares the results obtained with the CPD counterpart. Energy profiles for the multiple pathways were constructed with the help of Umbrella Sampling and same suitable collective variables used in Chapter Three. The pathways considered here were: (a) flipping out of 5’ dA before flipping of 3’ dA partner base, (b) flipping out of 3’ dA before that of 5’ dA, (c) inserion of BHD3-β hairpin in a lesion site where all bases were in extrahellicular state as well as (d) when all of them were intrahellicular. The pathways were constructed in the similar way as Chapter three’s pathways, i.e by generating metastable state and then using those states as starting points for next experiment in the pathway.

The results obtained from these pathways reveal that even in a 64PP lesioned DNA, there must be a cavity present for BHD3 β hairpin inertion to take place indicating that 64PP partner base flipping takes place before BHD3 β-hairpin inserts itself in the lesion site. Although 6-4PP containing complex promotes the insertion of BHD3-β hairpin with higher ease and flexibility compared to a CPD containing complex. Also, irrespective of the lesion present in the DNA 5’ dA partner base flips out before 3’ dA partner base as the even in presence of 64PP the pathway where 5’ dA flips out first is energetically more favourable.

The comparision of energy profiles obtained for association study on crystal structure and a structure where all the bases of lesion site are in extrahellicular state discloses that Rad4 is able to associate itself more strongly as the the recognition process moves forward.

In summary, irrespective of lesion type, 5' dA partner flips out before 3' dA partner base and partner base fliping preceds BHD3-β hairpin inserion. Although the energy barriers differ significantly in presence of different lesions. Also, during partner base flipping and hairpin insertion Rad4 binds more strongly to the lesioned DNA.

Future Work

In the current work a comparative analysis of partner base flipping energy could not be done due to different starting structures of flipping studies in CPD and 6-4PP. In enhanced sampling approaches, the precise energy difference between the flipping of partner bases in the presence of distinct lesions can be determined by utilizing identical starting structures. Also uncovering the effect of Rad4 on flipping of bases by performing Umbrella Sampling on a lesioned DNA in absence of binding protein can be done to find more insights into the role of Rad4 during partner base flipping. This couldn't be done in the current study due to the limitaions of distance based flipping collective variables. But the same collective variables can be used by making a dummy Center of Mass in place of CoMs of proteins used for distance calculations. Also, this study was performed with constrained computaional resources but with enough resources and 2-D or 3-D umbrella sampling, further insights into relationship between the above mentioned NER events can be obtained.

The study can also be expanded to study mismatch lesions instead of matched CPD and 64PP lesions.

Bibliography

[1] Rob Phillips, Jane Kondev, Julie Theriot, Hernan G. Garcia, and Nigel Orme. *Physical Biology of the Cell*. Garland Science, October 2012.

[2] Donald Ingber. Mechanobiology and diseases of mechanotransduction. *Annals of Medicine*, 35 (8):564–577, January 2003.

[3] Jamey D. Marth. A unified vision of the building blocks of life. *Nature Cell Biology*, 10(9): 1015–1015, September 2008.

[4] Harvey Lodish, Arnold Berk, Chris A Kaiser, Monty Krieger, Matthew P Scott, Anthony Bretscher, Hidde Ploegh, Paul Matsudaira, et al. *Molecular cell biology, 6th Edition*. Macmillan, 2008.

[5] Anastasia Khvorova, Aurélie Lescoute, Eric Westhof, and Sumedha D Jayasena. Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity. *Nature Structural & Molecular Biology*, 10(9):708, 2003.

[6] Colin R Hardwood and Anil Wipat. Bacterial protein synthesis. In *Molecular Medical Microbiology*, pages 321–338. Elsevier, 2002.

[7] Jeremy M Berg, John L Tymoczko, and Lubert Stryer. Biochemistry, 2012.

[8] Veit Hornung and Eicke Latz. Intracellular dna recognition. *Nature Reviews Immunology*, 10(2): 123, 2010.

[9] Mark E Tuckerman and Glenn J Martyna. *Understanding modern molecular dynamics: techniques and applications*. ACS Publications, 2000.

[10] Eric Paquet and Herna L Viktor. Molecular dynamics, monte carlo simulations, and langevin dynamics: a computational review. *BioMed Research International*, 2015, 2015.

[11] Gregory A Petsko and Dagmar Ringe. *Protein structure and function*. New Science Press, 2004.

[12] David Eisenberg, Edward M Marcotte, Ioannis Xenarios, and Todd O Yeates. Protein function in the post-genomic era. *Nature*, 405(6788):823, 2000.

[13] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, et al. International human genome sequencing consortium. initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[14] Loredana Lo Conte, Bart Ailey, Tim JP Hubbard, Steven E Brenner, Alexey G Murzin, and Cyrus Chothia. Scop: A structural classification of proteins database. *Nucleic Acids Research*, 28(1): 257–259, 2000.

[15] David Lee, Oliver Redfern, and Christine Orengo. Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*, 8(12):995, 2007.

[16] Sana Akbar, Kamal Raj Pardasani, and Nihar Ranjan Panda. PSO based neuro-fuzzy model for secondary structure prediction of protein. *Neural Processing Letters*, 53(6):4593–4612, August 2021.

[17] Soraya de Chadarevian. Portrait of a discovery: Watson, crick, and the double helix. *Isis*, 94(1): 90–105, 2003.

[18] Joshua Lederberg. The transformation of genetics by dna: an anniversary celebration of avery, macleod and mccarty (1944). *Genetics*, 136(2):423, 1994.

[19] Boris Magasanik, Ernst Vischer, Ruth Doniger, David Elson, and Erwin Chargaff. The separation and estimation of ribonucleotides in minute quantities. *The Journal of Biological Chemistry*, 186 (1):37–50, 1950.

[20] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171 (4356):737–738, 1953.

[21] Marshall W. Nirenberg and J. Heinrich Matthaei. The dependence of cell-free protein synthesis in e. coli upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences*, 47(10):1588–1602, October 1961.

[22] Christopher A Hunter. Sequence-dependent dna structure: the role of base stacking interactions. *The Journal of Biological Chemistry*, 230(3):1025–1054, 1993.

[23] Myron F Goodman. On the wagon-dna polymerase joins" h-bonds anonymous". *Nature Biotechnology*, 17(7):640, 1999.

[24] Leslie Pray. Discovery of dna structure and function: Watson and crick. *Nature Education*, 1(1): 100, 2008.

[25] Andrew Travers and Georgi Muskhelishvili. Dna structure and function. *The FEBS journal*, 282 (12):2279–2295, 2015.

[26] Karolin Luger, Armin W Mäder, Robin K Richmond, David F Sargent, and Timothy J Richmond. Crystal structure of the nucleosome core particle at 2.8 å resolution. *Nature*, 389(6648):251, 1997.

[27] M Michael Gromiha and R Nagarajan. Computational approaches for predicting the binding sites and understanding the recognition mechanism of protein–dna complexes. In *Advances in Protein Chemistry and Structural Biology*, volume 91, pages 65–99. Elsevier, 2013.

[28] Moyra Lawrence, Sylvain Daujat, and Robert Schneider. Lateral thinking: how histone modifications regulate gene expression. *Trends in Genetics*, 32(1):42–56, 2016.

[29] Maoxuan Lin and Jun tao Guo. New insights into protein–DNA binding specificity from hydrogen bond based comparative study. *Nucleic Acids Research*, 47(21):11103–11113, October 2019.

[30] Carsten Marr, Marcel Geertz, Marc-Thorsten Hütt, and Georgi Muskhelishvili. Dissecting the logical types of network control in gene expression profiles. *BMC Systems Biology*, 2(1):18, 2008.

[31] Matthew Slattery, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences*, 39(9):381–399, September 2014.

[32] Remo Rohs, Xiangshu Jin, Sean M. West, Rohit Joshi, Barry Honig, and Richard S. Mann. Origins of specificity in protein-DNA recognition. *Annual Review of Biochemistry*, 79(1):233–269, June 2010.

[33] Masashi Suzuki. A framework for the DNA–protein recognition code of the probe helix in transcription factors: the chemical and stereochemical rules. *Structure*, 2(4):317–326, April 1994.

[34] Nicholas M. Luscombe. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Research*, 29(13):2860–2874, July 2001.

[35] Yael Mandel-Gutfreund, Ora Schueler, and Hanah Margalit. Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: In search of common principles. *Journal of Molecular Biology*, 253(2):370–382, October 1995.

[36] Tsu-Pei Chiu, Satyanarayan Rao, Richard S. Mann, Barry Honig, and Remo Rohs. Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein–DNA binding. *Nucleic Acids Research*, 45(21):12565–12576, October 2017.

[37] Hassan Khesbak, Olesya Savchuk, Satoru Tsushima, and Karim Fahmy. The role of water h-bond imbalances in b-dna substate transitions and peptide recognition revealed by time-resolved ftir spectroscopy. *Journal of the American Chemical Society*, 133(15):5834–5842, 2011.

[38] Helmholtz Association of German Research Centres. Water molecules characterize the structure of dna genetic material, 2011.

[39] Nicholas M Luscombe, Susan E Austin, Helen M Berman, and Janet M Thornton. An overview of the structures of protein-dna complexes. *Genome Biology*, 1(1):reviews001.1, 2000.

[40] Yongping Pan, Chung-Jung Tsai, Buyong Ma, and Ruth Nussinov. Mechanisms of transcription factor selectivity. *Trends in Genetics*, 26(2):75–83, February 2010.

[41] George A. Jeffrey and Wolfram Saenger. *Hydrogen Bonding in Biological Structures*. Springer Berlin Heidelberg, 1991.

[42] Shandar Ahmad, Ozlem Keskin, Akinori Sarai, and Ruth Nussinov. Protein–DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. *Nucleic Acids Research*, 36(18):5922–5932, September 2008.

[43] Remo Rohs, Sean M. West, Alona Sosinsky, Peng Liu, Richard S. Mann, and Barry Honig. The role of DNA shape in protein–DNA recognition. *Nature*, 461(7268):1248–1253, October 2009.

[44] Gira Bhabha, Justin T Biel, and James S Fraser. Keep on moving: discovering and perturbing the conformational dynamics of enzymes. *Accounts of Chemical Research*, 48(2):423–430, 2014.

[45] Preeti Pandey, Sabeeha Hasnain, and Shandar Ahmad. Protein-dna interactions. *Encyclopedia of Bioinformatics and Computational Biology*, pages 142–154, 2019.

[46] Gregory R Bowman, Eric R Bolin, Kathryn M Hart, Brendan C Maguire, and Susan Marqusee. Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 112(9):2734–2739, 2015.

[47] Samuel Hertig, Naomi R Latorraca, and Ron O Dror. Revealing atomic-level mechanisms of protein allostery with molecular dynamics simulations. *PLOS Computational Biology*, 12(6): e1004746, 2016.

[48] Kenneth H. Kraemer. Sunlight and skin cancer:another link revealed. *Proceedings of the National Academy of Sciences of the United States of America*, 94(1):11–14, 1997. ISSN 0027-8424, 1091-6490.

[49] Martin G. Marinus. Dna mismatch repair. *EcoSal Plus*, 5(1), 2012.

[50] Joyce T Reardon and Aziz Sancar. Purification and characterization of escherichia coli and human nucleotide excision repair enzyme systems. *Methods in Enzymology*, 408:189–213, 2006.

[51] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9(9):646, 2002.

[52] Juan R Perilla, Boon Chong Goh, C Keith Cassidy, Bo Liu, Rafael C Bernardi, Till Rudack, Hang Yu, Zhe Wu, and Klaus Schulten. Molecular dynamics simulations of large macromolecular complexes. *Current Opinion in Structural Biology*, 31:64–74, 2015.

[53] Tom L Blundell and Louise N Johnson. *Protein crystallography*. Elsevier Science, 1976.

[54] Robert M Silverstein and G Clayton Bassler. Spectrometric identification of organic compounds. *Journal of Chemical Education*, 39(11):546, 1962.

[55] Hugo E Gottlieb, Vadim Kotlyar, and Abraham Nudelman. Nmr chemical shifts of common laboratory solvents as trace impurities. *The Journal of Organic Chemistry*, 62(21):7512–7515, 1997.

[56] Rieko Ishima and Dennis A Torchia. Protein dynamics from nmr. *Nature Structural Biology*, 7 (9):740, 2000.

[57] Helen Berman, Kim Henrick, Haruki Nakamura, and John L Markley. The worldwide protein data bank (wwpdb): ensuring a single, uniform archive of pdb data. *Nucleic Acids Research*, 35 (suppl_1):D301–D303, 2006.

[58] Samuel Flores, Nathaniel Echols, Duncan Milburn, Brandon Hespenheide, Kevin Keating, Jason Lu, Stephen Wells, Eric Z Yu, Michael Thorpe, and Mark Gerstein. The database of macromolecular motions: new features added at the decade mark. *Nucleic Acids Research*, 34(suppl_1): D296–D301, 2006.

[59] Michael P Allen et al. Introduction to molecular dynamics simulation. *Computational Soft Matter*, 23:1–28, 2004.

[60] Steve Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, 117(1):1–19, 1995.

[61] A Strey, Parallelrechner mit gemeinsamem Speicher, Parallelrechner mit verteiltem Speicher, and Parallelrechner mit virtuellem gemeinsamem Speicher. High performance computing. *International Encyclopedia of The Social & Behavioral Sciences*, pages 6693–6697, 2001.

[62] George S Almasi and Allan Gottlieb. *Highly parallel computing*. Menlo Park, CA (USA); Benjamin-Cummings Pub. Co., 1988.

[63] Romelia Salomon-Ferrer, Andreas W Götz, Duncan Poole, Scott Le Grand, and Ross C Walker. Routine microsecond molecular dynamics simulations with amber on gpus. *Journal of Chemical Theory and Computation*, 9(9):3878–3888, 2013.

[64] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12562–12566, 2002.

[65] Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

[66] John J. DiGiovanna and Kenneth H. Kraemer. Shining a light on xeroderma pigmentosum. *The Journal of Investigative Dermatology*, 132(3):785–796, 2012. ISSN 0022-202X.

[67] HaJeung Park, Kaijiang Zhang, Yingjie Ren, Sourena Nadji, Nanda Sinha, John-Stephen Taylor, and ChulHee Kang. Crystal structure of a DNA decamer containing a cis-syn thymine dimer. *Proceedings of the National Academy of Sciences*, 99(25):15965–15970, November 2002.

[68] Joon-Hwa Lee, Geum-Sook Hwang, and Byong-Seok Choi. Solution structure of a DNA decamer duplex containing the stable $3\prime$ t·g base pair of the pyrimidine(6–4)pyrimidone photoproduct [(6–4) adduct]: Implications for the highly specific $3\prime$ t $\rightarrow$ c transition of the (6–4) adduct. *Proceedings of the National Academy of Sciences*, 96(12):6632–6636, June 1999.

[69] Errol C Friedberg, Graham C Walker, Wolfram Siede, and Richard D Wood. *DNA repair and mutagenesis*. American Society for Microbiology Press, 2005.

[70] Nicholas E Geacintov and Suse Broyde. *Introduction and Perspectives on the Chemistry and Biology of DNA Damage*. Wiley Online Library, 2010.

[71] Jiri Lukas, Claudia Lukas, and Jiri Bartek. More than just a focus: The chromatin response to DNA damage and its role in genome integrity maintenance. *Nature Cell Biology*, 13(10):1161–1169, 2011. ISSN 1465-7392, 1476-4679.

[72] Dennis H. Oh and Graciela Spivak. Hereditary photodermatoses. In *Advances in Experimental Medicine and Biology*, pages 95–105. Springer New York, 2010.

[73] E. Cleaver James. UV damage, DNA repair and skin carcinogenesis. *Frontiers in Bioscience*, 7 (1-3):d1024, 2002.

[74] Gregory L. Verdine and Steven D. Bruner. How do DNA repair proteins locate damaged bases in the genome? *Chemistry & Biology*, 4(5):329–334, 1997. ISSN 1074-5521.

[75] Anjum Ansari and Serguei V. Kuznetsov. Dynamics and mechanism of DNA-bending proteins in binding site recognition. In *Biological and Medical Physics, Biomedical Engineering*, pages 107–142. Springer New York, 2010. ISBN 9780387928074, 9780387928081.

[76] Manas Kumar Sarangi, Viktoriya Zvoda, Molly Nelson Holte, Nicole A Becker, Justin P Peters, L James Maher, and Anjum Ansari. Evidence for a bind-then-bend mechanism for architectural DNA binding protein yNhp6A. *Nucleic Acids Research*, 47(6):2871–2883, 2019. ISSN 0305-1048, 1362-4962.

[77] Yogambigai Velmurugu, Paula Vivas, Mitchell Connolly, Serguei V Kuznetsov, Phoebe A Rice, and Anjum Ansari. Two-step interrogation then recognition of DNA binding site by integration host factor: An architectural DNA-bending protein. *Nucleic Acids Research*, 46(4):1741–1755, 2017. ISSN 0305-1048, 1362-4962.

[78] Elisa T. Zhang, Yuan He, Patricia Grob, Yick W. Fong, Eva Nogales, and Robert Tjian. Architecture of the human XPC DNA repair and stem cell coactivator complex. *Proceedings of the National Academy of Sciences of the United States of America*, 112(48):14817–14822, 2015. ISSN 0027-8424, 1091-6490.

[79] Lei Jia, Konstantin Kropachev, Shuang Ding, Bennett Van Houten, Nicholas E. Geacintov, and Suse Broyde. Exploring damage recognition models in prokaryotic nucleotide excision repair with a benzo[a]pyrene-derived lesion in UvrB. *Biochemistry (Moscow)*, 48(38):8948–8957, 2009. ISSN 0006-2960, 1520-4995.

[80] Debamita Paul, Hong Mu, Hong Zhao, Ouathek Ouerfelli, Philip D Jeffrey, Suse Broyde, and Jung-Hyun Min. Structure and mechanism of pyrimidine–pyrimidone (6-4) photoproduct recognition by the Rad4/XPC nucleotide excision repair complex. *Nucleic Acids Research*, 47(12): 6015–6028, 2019. ISSN 0305-1048.

[81] Sophie E. Polo and Stephen P. Jackson. Dynamics of DNA damage response proteins at DNA breaks: A focus on protein modifications. *Genes & Development*, 25(5):409–433, 2011. ISSN 0890-9369.

[82] Alberto Ciccia and Stephen J. Elledge. The DNA damage response: Making it safe to play with knives. *Molecular Cell*, 40(2):179–204, 2010. ISSN 1097-2765.

[83] Jung-Hyun Min and Nikola P. Pavletich. Recognition of DNA damage by the rad4 nucleotide excision repair protein. *Nature*, 449(7162):570–575, September 2007.

[84] Debamita Paul, Hong Mu, Hong Zhao, Ouathek Ouerfelli, Philip D Jeffrey, Suse Broyde, and Jung-Hyun Min. Structure and mechanism of pyrimidine–pyrimidone (6-4) photoproduct recog-

nition by the rad4/XPC nucleotide excision repair complex. *Nucleic Acids Research*, 47(12): 6015–6028, May 2019.

[85] Vivek Anantharaman, Eugene V Koonin, and L Aravind. Peptide-n-glycanases and dna repair proteins, xp-c/rad4, are, respectively, active and inactivated enzymes sharing a common transglutaminase fold. *Human molecular genetics*, 10(16):1627–1630, 2001.

[86] Jung-Hyun Min and Nikola P. Pavletich. Crystal structure of the rad4-rad23 complex, October 2007.

[87] Xuejing Chen, Yogambigai Velmurugu, Guanqun Zheng, Beomseok Park, Yoonjung Shim, Youngchang Kim, Lili Liu, Bennett Van Houten, Chuan He, Anjum Ansari, and Jung-Hyun Min. Kinetic gating mechanism of DNA damage recognition by Rad4/XPC. *Nature Communications*, 6(1):5849, 2015. ISSN 2041-1723.

[88] Jung-Hyun Min and Nikola P. Pavletich. Recognition of DNA damage by the rad4 nucleotide excision repair protein. *Nature*, 449(7162):570–575, 2007. ISSN 0028-0836, 1476-4687.

[89] Yang Liu, Dara Reeves, Konstantin Kropachev, Yuqin Cai, Shuang Ding, Marina Kolbanovskiy, Alexander Kolbanovskiy, Judith L. Bolton, Suse Broyde, Bennett Van Houten, and Nicholas E. Geacintov. Probing for DNA damage with β-hairpins: Similarities in incision efficiencies of bulky DNA adducts by prokaryotic and human nucleotide excision repair systems in vitro. *DNA Repair*, 10(7):684–696, 2011. ISSN 1568-7864.

[90] Hong Mu, Nicholas E. Geacintov, Yingkai Zhang, and Suse Broyde. Recognition of damaged DNA for nucleotide excision repair: A correlated motion mechanism with a mismatched cis-syn thymine dimer lesion. *Biochemistry (Moscow)*, 54(34):5263–5267, 2015. ISSN 0006-2960, 1520-4995.

[91] Hong Mu, Nicholas E Geacintov, Jung-Hyun Min, Yingkai Zhang, and Suse Broyde. Nucleotide excision repair lesion-recognition protein rad4 captures a pre-flipped partner base in a benzo [a] pyrene-derived DNA lesion: How structure impacts the binding pathway. *Chemical Research in Toxicology*, 30(6):1344–1354, 2017.

[92] Berni Julian Alder and Thomas Everett Wainwright. Phase transition for a hard sphere system. *The Journal of Chemical Physics*, 27(5):1208–1209, 1957.

[93] Scott A Hollingsworth and Ron O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6): 1129–1143, 2018.

[94] Andrew R Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.

[95] Ken Dill and Sarina Bromberg. *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience*. Garland Science, 2012.

[96] Simon W de Leeuw, John W Perram, and Henrik G Petersen. Hamilton's equations for constrained dynamical systems. *Journal of Statistical Physics*, 61(5-6):1203–1222, 1990.

[97] J Willard Gibbs. Elementary principles in statistical mechanics, developed with especial reference to the rational foundation of thermodynamics, 1902. *Nachdruck*, 1960.

[98] Donald G Truhlar, Bruce C Garrett, and Stephen J Klippenstein. Current status of transition-state theory. *The Journal of Physical Chemistry A*, 100(31):12771–12800, 1996.

[99] Tamar Schlick. *Molecular modeling and simulation: an interdisciplinary guide: an interdisciplinary guide*, volume 21. Springer Science & Business Media, 2010.

[100] William C Swope, Hans C Andersen, Peter H Berens, and Kent R Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649, 1982.

[101] Hans Martin Senn and Walter Thiel. Qm/mm methods for biomolecular systems. *Angewandte Chemie International Edition*, 48(7):1198–1229, 2009.

[102] Rodrigo Galindo-Murillo, James C Robertson, Marie Zgarbova, Jiri Sponer, Michal Otyepka, Petr Jurečka, and Thomas E Cheatham III. Assessing the current state of amber force field modifications for dna. *Journal of Chemical Theory and Computation*, 12(8):4114–4127, 2016.

[103] Geoffrey C Maitland, Maurice Rigby, E Brian Smith, and William A Wakeham. *Intermolecular forces: their origin and determination*, volume 3. Clarendon Press Oxford, 1981.

[104] Anthony J Stone. *The theory of intermolecular forces*. Oxford, 1996.

[105] Michiel Sprik. Effective pair potentials and beyond. In *Computer simulation in chemical physics*, pages 211–259. Springer, 1993.

[106] Hasitha Muthumala Waidyasooriya, Masanori Hariyama, and Kota Kasahara. An FPGA accelerator for molecular dynamics simulation using OpenCL. *International Journal of Networked and Distributed Computing*, 5(1):52, 2017.

[107] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 2015.

[108] J. E. Jones. On the determination of molecular fields. II. from the equation of state of a gas. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 106(738): 463–477, October 1924.

[109] Jiapu Zhang. Canonical dual theory applied to a lennard-jones potential minimization problem. *Practical Global Optimization Computing Methods in Molecular Modelling - for Atomic-resolution Structures of Amyloid Fibrils*, pages 94–114, 2011.

[110] Magnus Rudolph Hestenes and Eduard Stiefel. *Methods of conjugate gradients for solving linear systems*, volume 49. NBS Washington, DC, 1952.

[111] Tomas Hansson, Chris Oostenbrink, and WilfredF van Gunsteren. Molecular dynamics simulations. *Current Opinion in Structural Biology*, 12(2):190–196, April 2002.

[112] Roland Stote, Annick Dejaegere, Dmitry Kuznetsov, and Laurent Falquet. Theory of molecular dynamics simulations. URL `https://www.ch.embnet.org/MD_tutorial/pages/MD.Part1.html`.

[113] Loup Verlet. Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Physical Review*, 159(1):98, 1967.

[114] Loup Verlet. Computer "experiments" on classical fluids. ii. Equilibrium correlation functions. *Physical Review*, 165(1):201, 1968.

[115] Herman JC Berendsen, JPM van Postma, Wilfred F van Gunsteren, ARHJ DiNola, and JR Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8): 3684–3690, 1984.

[116] Glenn M Torrie and John P Valleau. Monte carlo free energy estimates using non-boltzmann sampling: Application to the sub-critical lennard-jones fluid. *Chemical Physics Letters*, 28(4): 578–581, 1974.

[117] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992. ISSN 0192-8651, 1096-987X.

[118] Marc Souaille and Benoıt Roux. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Computer Physics Communications*, 135 (1):40–57, 2001.

[119] Johannes Kästner and Walter Thiel. Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method:"umbrella integration". *The Journal of Chemical Physics*, 123(14):144104, 2005.

[120] Ratna S. Katiyar and Prateek K. Jha. Molecular simulations in drug delivery: Opportunities and challenges. *WIREs Computational Molecular Science*, 8(4), February 2018.

[121] Donald Ervin Knuth. *The art of computer programming*, volume 3. Pearson Education, 1997.

[122] Roger W Hockney and James W Eastwood. *Computer simulation using particles*. CRC Press, 1988.

[123] Wan-Qing Li, Tang Ying, Wan Jian, and Dong-Jin Yu. Comparison research on the neighbor list algorithms: Verlet table and linked-cell. *Computer Physics Communications*, 181(10):1682–1686, October 2010.

[124] Tom Darden, Lalith Perera, Leping Li, and Lee Pedersen. New tricks for modelers from the crystallography toolkit: the particle mesh ewald algorithm and its use in nucleic acid simulations. *Structure*, 7(3):R55–R60, 1999.

[125] Michele Di Pierro, Ron Elber, and Benedict Leimkuhler. A stochastic algorithm for the isobaric–isothermal ensemble with ewald summations for all long range forces. *Journal of Chemical Theory and Computation*, 11(12):5624–5637, 2015.

[126] Paul P Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. *Annals of Physics*, 369(3):253–287, 1921.

[127] Jiri Kolafa and John W Perram. Cutoff errors in the ewald summation formulae for point charge systems. *Molecular Simulation*, 9(5):351–368, 1992.

[128] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An n· log (n) method for ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.

[129] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman JC Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977.

[130] Giovanni Ciccotti and Jean-Paul Ryckaert. Molecular dynamics simulation of rigid molecules. *Computer Physics Reports*, 4(6):346–392, 1986.

[131] Hans C Andersen. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *Journal of Computational Physics*, 52(1):24–34, 1983.

[132] Alexander Hollaender. Effect of long ultraviolet and short visible radiation (3500 to 4900Å) on escherichia coli. *Journal of Bacteriology*, 46(6):531–541, 1943.

[133] Hermann Joseph Muller. Artificial transmutation of the gene. *Science*, 66(1699):84–87, 1927.

[134] Edgar Altenburg. The artificial production of mutations by ultra-violet light. *The American Naturalist*, 68(719):491–507, 1934.

[135] EC Freidberg. Correcting the blueprint of life. *An Historical Accounting of the Discovery DNA Repairing Mechanisms*, 1997.

[136] Errol C Friedberg. A brief history of the DNA repair field. *Cell Research*, 18(1):3–7, December 2007.

[137] R. B. Setlow. The wavelengths in sunlight effective in producing skin cancer: A theoretical analysis. *Proceedings of the National Academy of Sciences*, 71(9):3363–3366, September 1974.

[138] Luís F.Z. Batista, Bernd Kaina, Rogério Meneghini, and Carlos F.M. Menck. How DNA lesions are turned into powerful killing structures: Insights from UV-induced apoptosis. *Mutation Research/Reviews in Mutation Research*, 681(2-3):197–208, March 2009.

[139] Richard B. Setlow and William L. Carrier. The disappearance of thymine dimers from dna: an error-correcting mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 51(2):226, 1964.

[140] Richard P Boyce and Paul Howard-Flanders. Release of ultraviolet light-induced thymine dimers from dna in e. coli k-12. *Proceedings of the National Academy of Sciences of the United States of America*, 51(2):293, 1964.

[141] David Pettijohn and Philip Hanawalt. Evidence for repair-replication of ultraviolet damaged dna in bacteria. *Journal of Molecular Biology*, 9(2):395–410, 1964.

[142] Martin T. Hess, Urs Schwitter, Mario Petretta, Bernd Giese, and Hanspeter Naegeli. Bipartite substrate discrimination by human nucleotide excision repair. *Proceedings of the National Academy of Sciences*, 94(13):6664–6669, June 1997.

[143] Kaoru Sugasawa, Tomoko Okamoto, Yuichiro Shimizu, Chikahide Masutani, Shigenori Iwai, and Fumio Hanaoka. A multistep damage recognition mechanism for global genomic nucleotide excision repair. *Genes & Development*, 15(5):507–521, March 2001.

[144] Kaoru Sugasawa, Jessica M.Y Ng, Chikahide Masutani, Shigenori Iwai, Peter J van der Spek, André P.M Eker, Fumio Hanaoka, Dirk Bootsma, and Jan H.J Hoeijmakers. Xeroderma pigmentosum group c protein complex is the initiator of global genome nucleotide excision repair. *Molecular Cell*, 2(2):223–232, August 1998.

[145] Marcel Volker, Martijn J Moné, Parimal Karmakar, Anneke van Hoffen, Wouter Schul, Wim Vermeulen, Jan H.J Hoeijmakers, Roel van Driel, Albert A van Zeeland, and Leon H.F Mullenders. Sequential assembly of the nucleotide excision repair factors in vivo. *Molecular Cell*, 8(1): 213–224, July 2001.

[146] Thilo Riedl, Fumio Hanaoka, and Jean-Marc Egly. The comings and goings of nucleotide excision repair factors on damaged DNA. *The EMBO Journal*, 22(19):5293–5303, October 2003.

[147] Thierry Nouspikel. DNA repair in mammalian cells. *Cellular and Molecular Life Sciences*, 66 (6):994–1009, January 2009.

[148] Ulrike Camenisch, Daniel Träutlein, Flurina C Clement, Jia Fei, Alfred Leitenstorfer, Elisa Ferrando-May, and Hanspeter Naegeli. Two-stage dynamic dna quality check by xeroderma pigmentosum group c protein. *The EMBO Journal*, 28(16):2387–2399, 2009.

[149] Muwen Kong, Lili Liu, Xuejing Chen, Katherine I. Driscoll, Peng Mao, Stefanie Böhm, Neil M. Kad, Simon C. Watkins, Kara A. Bernstein, John J. Wyrick, Jung-Hyun Min, and Bennett Van Houten. Single-molecule imaging reveals that rad4 employs a dynamic dna damage recognition process. *Molecular Cell*, 64(2):376–387, 2016.

[150] Stephen E Halford and John F Marko. How do site-specific dna-binding proteins find their targets? *Nucleic Acids Research*, 32(10):3040–3052, 2004.

[151] Otto G Berg, Robert B Winter, and Peter H Von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. models and theory. *Biochemistry*, 20(24):6929–6948, 1981.

[152] Paul C Blainey, Antoine M van Oijen, Anirban Banerjee, Gregory L Verdine, and X Sunney Xie. A base-excision dna-repair protein finds intrahelical lesion bases by fast sliding in contact with dna. *Proceedings of the National Academy of Sciences*, 103(15):5752–5757, 2006.

[153] Yogambigai Velmurugu, Xuejing Chen, Phillip Slogoff Sevilla, Jung-Hyun Min, and Anjum Ansari. Twist-open mechanism of dna damage recognition by the rad4/xpc nucleotide excision repair complex. *Proceedings of the National Academy of Sciences*, 113(16):E2296–E2305, 2016.

[154] Patrick Calsou and Bernard Salles. Properties of damage-dependent dna incision by nucleotide excision repair in human cell-free extracts. *Nucleic Acids Research*, 22(23):4937–4942, 1994.

[155] Elizabeth Evans, Jane Fellows, Arnold Coffer, and Richard D Wood. Open complex formation around a lesion during nucleotide excision repair provides a structure for cleavage by human xpg protein. *The EMBO Journal*, 16(3):625–638, 1997.

[156] Elizabeth Evans, Jonathan G Moggs, Jae R Hwang, Jean-Marc Egly, and Richard D Wood. Mechanism of open complex and dual incision formation by human nucleotide excision repair factors. *The EMBO Journal*, 16(21):6559–6573, 1997.

[157] David Mu, Mitsuo Wakasugi, David S Hsu, and Aziz Sancar. Characterization of reaction intermediates of human excision repair nuclease. *Journal of Biological Chemistry*, 272(46):28971–28979, 1997.

[158] Fr'ed'eric Coin, Valentyn Oksenych, and Jean-Marc Egly. Distinct roles for the xpb/p52 and xpd/p44 subcomplexes of tfiih in damaged dna opening during nucleotide excision repair. *Molecular Cell*, 26(2):245–256, 2007.

[159] John Bradsher, Frederic Coin, and Jean-Marc Egly. Distinct roles for the helicases of tfiih in transcript initiation and promoter escape. *Journal of Biological Chemistry*, 275(4):2532–2538, 2000.

[160] G Sebastiaan Winkler, Sofia J Araújo, Ulrike Fiedler, Wim Vermeulen, Frederic Coin, Jean-Marc Egly, Jan HJ Hoeijmakers, Richard D Wood, H Th Marc Timmers, and Geert Weeda. Tfiih with inactive xpd helicase functions in transcription initiation but is defective in dna repair. *Journal of Biological Chemistry*, 275(6):4258–4266, 2000.

[161] Li Fan, Jill O Fuss, Quen J Cheng, Andrew S Arvai, Michal Hammel, Victoria A Roberts, Priscilla K Cooper, and John A Tainer. Xpd helicase structures and activities: insights into the cancer and aging phenotypes from xpd mutations. *Cell*, 133(5):789–800, 2008.

[162] Huanting Liu, Jana Rudolf, Kenneth A Johnson, Stephen A McMahon, Muse Oke, Lester Carter, Anne-Marie McRobbie, Sara E Brown, James H Naismith, and Malcolm F White. Structure of the dna repair helicase xpd. *Cell*, 133(5):801–812, 2008.

[163] Stefanie C Wolski, Jochen Kuper, Petra Hänzelmann, James J Truglio, Deborah L Croteau, Bennett Van Houten, and Caroline Kisker. Crystal structure of the fes cluster–containing nucleotide excision repair helicase xpd. *PLoS Biology*, 6(6):e149, 2008.

[164] Nadine Mathieu, Nina Kaczmarek, and Hanspeter Naegeli. Strand-and site-specific dna lesion demarcation by the xeroderma pigmentosum group d helicase. *Proceedings of the National Academy of Sciences*, 107(41):17545–17550, 2010.

[165] Jochen Kuper, Stefanie C Wolski, Gudrun Michels, and Caroline Kisker. Functional and structural studies of the nucleotide excision repair helicase xpd suggest a polarity for dna translocation. *The EMBO Journal*, 31(2):494–502, 2012.

[166] Robert A Pugh, Colin G Wu, and Maria Spies. Regulation of translocation polarity by helicase domain 1 in sf2b helicases. *The EMBO Journal*, 31(2):503–514, 2012.

[167] Juch-Chin Huang, Daniel L Svoboda, Joyce T Reardon, and Aziz Sancar. Human nucleotide excision nuclease removes thymine dimers from dna by incising the 22nd phosphodiester bond 5'and the 6th phosphodiester bond 3'to the photodimer. *Proceedings of the National Academy of Sciences*, 89(8):3664–3668, 1992.

[168] Jonathan G Moggs, Kevin J Yarema, John M Essigmann, and Richard D Wood. Analysis of incision sites produced by human cell extracts and purified proteins during nucleotide excision repair of a 1, 3-intrastrand d(gptpg)-cisplatin adduct (*). *Journal of Biological Chemistry*, 271 (12):7177–7186, 1996.

[169] Anne O'Donovan, Adelina A Davies, Jonathan G Moggs, Stephen C West, and Richard D Wood. Xpg endonuclease makes the 3' incision in human dna nucleotide excision repair. *Nature*, 371 (6496):432–435, 1994.

[170] Angelos Constantinou, Daniela Gunz, Elizabeth Evans, Philippe Lalle, Paul A Bates, Richard D Wood, and Stuart G Clarkson. Conserved residues of human xpg protein important for nuclease activity and function in nucleotide excision repair. *Journal of Biological Chemistry*, 274(9): 5637–5648, 1999.

[171] Odilia Popanda and Heinz Walter Thielmann. The function of dna polymerases in dna repair synthesis of ultraviolet-irradiated human fibroblasts. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, 1129(2):155–160, 1992.

[172] Silvano Nocentini. Rejoining kinetics of DNA single- and double-strand breaks in normal and DNA ligase-deficient cells after exposure to ultraviolet c and gamma radiation: An evaluation of ligating activities involved in different DNA repair processes. *Radiation Research*, 151(4):423, April 1999.

[173] Intisar Husain, Jack Griffith, and Aziz Sancar. Thymine dimers bend DNA. *Proceedings of the National Academy of Sciences*, 85(8):2558–2562, April 1988.

[174] David A. Pearlman, Stephen R. Holbrook, David H. Pirkle, and Sung-Hou Kim. Molecular models for DNA damaged by photoreaction. *Science*, 227(4692):1304–1308, March 1985.

[175] Renata MA Costa, Vanessa Chiganças, Rodrigo da Silva Galhardo, Helotonio Carvalho, and Carlos FM Menck. The eukaryotic nucleotide excision repair pathway. *Biochimie*, 85(11):1083–1099, 2003.

[176] Luca Proietti De Santis, Claudia Lorenti Garcia, Adayabalam S Balajee, Paolo Latini, Pietro Pichierri, Osamu Nikaido, Miria Stefanini, and Fabrizio Palitti. Transcription coupled repair efficiency determines the cell cycle progression and apoptosis after uv exposure in hamster cells. *DNA Repair*, 1(3):209–223, 2002.

[177] James M Ford. Regulation of dna damage recognition and nucleotide excision repair: another role for p53. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 577 (1-2):195–202, 2005.

[178] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[179] Andrej Šali and Tom L Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, 1993.

[180] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. UCSF chimera visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, 2004.

[181] Thomas J. Macke and David A. Case. Modeling unusual nucleic acid structures. In *ACS Symposium Series*, pages 379–393. American Chemical Society, July 1997.

[182] David Case, Robin M. Betz, D.S. Cerutti, Thomas Cheatham, Thomas Darden, Robert Duke, T.J. Giese, Holger Gohlke, Andreas Götz, Nadine Homeyer, Saeed Izadi, Pawel Janowski, J Kaus, Andriy Kovalenko, Tai-Sung Lee, S LeGrand, P Li, C Lin, Tyler Luchko, and Peter A. Kollman. Amber 18, university of california, san francisco. *University of California, San Francisco*, pages 2246–2257, 2018.

[183] Romelia Salomon-Ferrer, Andreas W. Götz, Duncan Poole, Scott Le Grand, and Ross C. Walker. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. explicit solvent particle mesh ewald. *Journal of Chemical Theory and Computation*, 9(9):3878–3888, August 2013.

[184] Ivan Ivani, Pablo D Dans, Agnes Noy, Alberto Pérez, Ignacio Faustino, Adam Hospital, Jürgen Walther, Pau Andrio, Ramon Goñi, and Alexandra Balaceanu. Parmbsc1: a refined force field for dna simulations. *Nature Methods*, 13(1):55, 2016.

[185] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.

[186] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9): 1157–1174, 2004.

[187] Richard J Loncharich, Bernard R Brooks, and Richard W Pastor. Langevin dynamics of peptides: The frictional dependence of isomerization rates of n-acetylalanyl-n′-methylamide. *Biopolymers: Original Research on Biomolecules*, 32(5):523–535, 1992.

[188] Hong Mu, Nicholas E. Geacintov, Yingkai Zhang, and Suse Broyde. Recognition of damaged DNA for nucleotide excision repair: A correlated motion mechanism with a mismatchedcis-synThymine dimer lesion. *Biochemistry*, 54(34):5263–5267, August 2015.

[189] Jung-Hyun Min and Phil D. Jeffrey. Crystal structure of rad4-rad23 bound to a 6-4 photoproduct UV lesion, February 2019.

[190] Abhinandan Panigrahi, Hemanth Vemuri, Madhur Aggarwal, Kartheek Pitta, and Marimuthu Krishnan. Sequence specificity, energetics and mechanism of mismatch recognition by DNA damage sensing protein rad4/XPC. *Nucleic Acids Research*, 48(5):2246–2257, February 2020.

[191] Joon-Hwa Lee. NMR structure of the DNA decamer duplex containing double t.g mismatches of cis-syn cyclobutane pyrimidine dimer: implications for DNA damage recognition by the XPC-hHR23b complex. *Nucleic Acids Research*, 32(8):2474–2481, April 2004.

[192] Christopher I. Bayly, Piotr Cieplak, Wendy Cornell, and Peter A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry*, 97(40):10269–10280, October 1993.

[193] Zara Molphy, Creina Slator, Chryssostomos Chatgilialoglu, and Andrew Kellett. Dna oxidation profiles of copper phenanthrene chemical nucleases. *Frontiers in Chemistry*, 3:66–74, 10 2015.

[194] Jong-Ki Kim, Dinshaw Patel, and Byong-Seok Choi. Contrasting structural impacts induced by cis-syn cyclobutane dimer and 6–4 adduct in dna duplex decamers: Implication in mutagenesis and repair activity. *Photochemistry and Photobiology*, 62(1):44–50, July 1995.