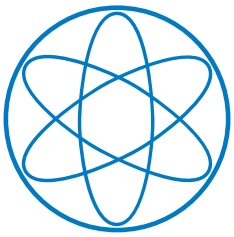
Physik Department



RECOGNITION OF UV-INDUCED DNA DAMAGE

A Molecular Dynamics Analysis

Dissertation
von

ALEXANDER KNIPS

Technische Universität München

Fakultät der Physik

Lehrstuhl für Theoretische Biophysik T38 - Molekulardynamik

Recognition of UV-induced DNA damage
A Molecular Dynamics Analysis

ALEXANDER KNIPS

Vollständiger Abdruck der von der Fakultät der Physik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Friedrich C. Simmel
Prüfer der Dissertation:	1. Univ.-Prof. Dr. Martin Zacharias
	2. Univ.-Prof. Dr. Iris Antes

Die Dissertation wurde am 09.11.2015 bei der Technischen Universität München eingereicht und durch die Fakultät der Physik am 16.12.2015 angenommen.

ABSTRACT

Damage of the Deoxyribonucleic Acid (DNA) is dangerous to all forms of life. It can lead to mutations of the DNA and sometimes to cancer. The lesions in DNA can be caused by many different processes. One of these is the absorption of Ultra-Violet (UV) light which can lead to direct damage of the DNA. The most common form of UV-induced damage is the Cyclobutane Pyrimidine Dimer (CPD) lesion.

For a long time it has been a question how repair enzymes effectively detect and mend those types of damages. In bacteria such as Escherichia coli (E. coli) photolyases can efficiently reverse the dimer formation employing a light-driven reaction after looping out the CPD damaged bases into the enzyme active site. The exact mechanism of how the repair enzyme identifies a damaged site within a large surplus of undamaged DNA is not fully understood yet.

One explanation is based on the structural effects of the CPD damage on DNA. In particular, the CPD damage may alter the DNA structure and dynamics already in the absence of the repair enzyme, which can ease the initial binding of a photolyase repair enzyme.

To characterize the effect of a CPD damage, extensive comparative Molecular Dynamics (MD) simulations on duplex DNA with central regular or CPD damaged nucleotides were performed, supplemented by simulations of the DNA-photolyase complex. Although no spontaneous flipping out transitions of the damaged bases were observed, the simulations showed significant differences in the conformational states of regular and CPD damaged DNA.

The unrestrained simulations were joined by Umbrella Sampling MD simulations along the flipping reaction coordinate. The flipping transition was analyzed for native and damaged DNA, studied in bulk water and in the complex with the photolyase enzyme. A tendency towards favorable flipping of damaged bases was found. In particular the flipping happens through the major groove of the DNA.

To understand the complete, complex repair and recognition mechanism in detail, Umbrella Sampling (US) simulations of the structural change during binding were performed. In particular, the transition was studied along a two-dimensional RMSD reaction coordinate employing Hamiltonian Replica Exchange Molecular Dynamics (HREMD) for better sampling. Hence, it could be determined that damaged DNA more easily undergoes the transition

into the conformation necessary for a close binding with the repair enzyme.

Taking the results of all these simulations together, a complete picture of the recognition and repair of CPD lesion in DNA was obtained. It was shown that the structural differences of damaged DNA in comparison to undamaged DNA are primarily responsible for the recognition mechanism and that the local interaction with the photolyase protein facilitates the flipping response of damaged DNA in contact with the repair protein.

ZUSAMMENFASSUNG

DNA Schäden haben großes Schadenspotential für alle Arten des Lebens. Sie können zu Mutationen führen, von denen manche letztendlich zu Krebs führen können. Die Beschädigungen der DNA können durch multiple Faktoren verursacht werden. Einer dieser Faktoren ist die Adsorption von ultra-violettem Licht, welche DNA direkt und indirekt beschädigen kann. Der Cyclobutane Pyrimidine Dimer (CPD)-Schaden ist eine direkte Beschädigung der DNA, wobei zwei nebeneinanderliegende Pyrimidine zu einem Dimer verbunden werden. CPD-Schäden sind eine der häufigsten durch UV-Licht hervorgerufenen Formen der DNA Beschädigung.

Seit längerem wird diskutiert, inwiefern Reparatur-Enzyme CPD-Schäden so effizient finden und reparieren können, wie es beobachtet wird. Beispielsweise können in Bakterien wie Escherichia coli (E. coli) Photolyasen CPD-Läsionen in einer großen Anzahl von unbeschädigten DNA Basen effizient auffinden und unter Zuhilfenahme von blauem Licht reparieren. Jedoch ist der Mechanismus der Erkennung der beschädigten Basen noch nicht vollständig verstanden.

Die strukturellen Einflüsse der CPD-Schäden auf die DNA können hierbei eine große Rolle spielen. Durch Änderung ihrer Struktur und Dynamik könnte beschädigte DNA leichter an das Reparatur-Protein binden. Um diese Effekte zu untersuchen, wurden im Rahmen dieser Doktorarbeit umfangreiche Simulationen der DNA durchgeführt. Diese wurden verglichen mit Simulationen von unbeschädigter DNA sowie Simulationen der DNA im Komplex mit dem Reparatur-Enzym. In keiner der Simulationen wurde ein spontaner Flip der beschädigten nebeneinanderliegenden Basen von der intrahelikalen in die extrahelikale Position beobachtet. Dennoch konnte gezeigt werden, dass erhebliche strukturelle Unterschiede zwischen beschädigter und nativer DNA bestehen. Um den Flip-Mechanismus genauer zu verstehen wurden US-MD Simulationen benutzt, wobei die DNA in verschiedenen Konfigurationen und im Komplex mit dem Reparatur-Enzym untersucht wurde. Bei beschädigten Basen und in Anwesenheit des Reparatur-Enzyms wurde eine Tendenz zu einem einfacheren Flip-Verhalten beobachtet. Im Speziellen dreht sich der CPD-Dimer über die große "DNA-Furche" von der intrahelikalen in die extrahelikale Position.

Da dies den Erkennungsmechanismus von beschädigter DNA durch Photolyasen aber nicht komplett erklären konnte, wurde das Verhalten der DNA im Erkennungs-Komplex genauer untersucht. Die DNA muss dabei ihre globale Struktur verändern, um nah mit

dem Reparatur-Protein in Kontakt zu kommen. US Simulationen entlang der RMSD Koordinate zeigten, dass diese strukturelle Änderung, die für eine enge Bindung zwischen DNA und Photolyase notwendig ist, leichter eingenommen werden kann, wenn die DNA beschädigt ist. Die Methode des Umbrella Samplings wurde hierzu auf zwei RMSD Koordinaten für eine genauere Betrachtung erweitert.

Somit konnte letztendlich gezeigt werden, dass die strukturellen Unterschiede im Bereich von ca. 10 Basenpaaren zwischen beschädigter und unbeschädigter DNA für die Erkennung von CPD-Schäden unter einer Vielzahl von unbeschädigten Basen verantwortlich sind. Die Reparatur und das Drehen der beschädigten Basen an sich sind weitestgehend durch die Interaktionen des Reparatur-Enzyms mit der beschädigten DNA bestimmt.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

Chapter 6: Influence of a cis,syn-cyclobutane pyrimidine dimer damage on DNA conformation studied by molecular dynamics simulations in Biopolymers (2014) [89]

Chapter 7: Submitted for publication in shorter form.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

6-4PP	6-4 Photo Product
8-oxoG	8-oxo Guanine
A	Adenine
A-DNA	A-form DNA
AMBER	Assisted Model Building with Energy Refinement
AP site	Abasic site
ARG	Arginine
BAR	Bennett Acceptance Ratio
B-DNA	B-form DNA
BER	Base Excision Repair
C	Cytosine
CPD	Cyclobutane Pyrimidine Dimer
CPU	Central Processing Unit
DFT	Density Functional Theory
DNA	Deoxyribonucleic Acid
D-RMSD	Distance-RMSD
DSB	Double Strand Break
ds-DNA	double stranded DNA
E. coli	Escherichia coli
ERCC1-XPF	Excision repair cross-complementation Group - Xeroderma pigmentosum, complementation Group F
ext.	extra-helical
FAD	Flavin Adenine Dinucleotide
FADH	Flavin Adenine Dinucleotide free radical
FEP	Free Energy Perturbation
FFT	Fast Fourier Transform

FRET	Förster Resonance Energy Transfer or Fluorescence Resonance Energy Transfer
GB	Generalized Born
GGR	Global Genome Repair
G	Guanine
GPU	Graphics Processor Unit
H-bond	Hydrogen bond
HhaI	HhaI DNA methyltransferase
HREMD	Hamiltonian Replica Exchange Molecular Dynamics
HR	Homologous Recombination
incrd.	increased
int.	intra-helical
IR	Infra-red Radiation
LP-BER	Long-Patch BER
MC	Monte Carlo
MD	Molecular Dynamics
MFPT	Mean First Passage Time
M. HgaI	Modification methylase HgaI
MMR	MisMatch Repair
MTHF	Methenyl Tetra Hydro Folate
MutM	Formamidopyrimidine DNA glycosylase
NAB	Nucleic Acid Builder
ncc.	not complete coverage
NER	Nucleotide Excision Repair
NHEJ	Non-Homologous End Joining
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
PCB	Periodic Boundary Conditions
PCD	Programmed Cell Death

PDB	Protein DataBase
PMEMD	Particle Mesh Ewald Molecular Dynamics
PME	Particle Mesh Ewald
PMF	Potential of mean force
PRO	Proline
QM/MD	Quantum Mechanics/Molecular Dynamics hybrid
RATTLE	Algorithm for Rigid Water Models
REMD	Replica Exchange Molecular Dynamics
RMSD	Root-mean-square deviation
RNA	Ribonucleic Acid
ROS	Reactive Oxygen Species
S. cerevisae	Saccharomyces cerevisiae
SOS response	Save Our Soul response
SP-BER	Short-Patch BER
SSB	Single Strand Break
TCR	Transcription Coupled Repair
TFIIH	Transcription factor II Human
TI	Thermodynamic Integration
TRP	Tryptophan
T	Thymine
TT	Thymine Thymine adjacent pair
US	Umbrella Sampling
U	Uracil
UVR	Ultra-Violet Radiation
UV	Ultra-Violet
WC-pair	Watson-Crick pair
WC	Watson-Crick
WHAM	Weighted Histogram Analysis Method
w.	with

XPC-hHR23B Xeroderma Pigmentosum Complementation group Human Rad23 Homolog complex

XPC-RAD23B Xeroderma Pigmentosum Complementation group yeast Rad23 complex

XPG Xeroderma Pigmentosum, complementation group G

Z-DNA Z-form DNA

Part I

INTRODUCTION, THEORY, MODEL, IMPORTANT REFERENCES RESULTS, AND THE METHOD OF SETUP

In this chapter an outline of the biological problem at hand will be given. The basics about DNA repair will be explained and multiple hypotheses studied during my PhD will be lain out here. The chapter following the introduction will give an overview about the theory needed to understand the methods and results of the thesis. Important recent MD simulation studies on the topic will be summarized.

1 INTRODUCTION

1.1 NUCLEIC ACIDS AND DNA

The most important macromolecules essential for life are proteins, carbohydrates and nucleic acids. Nucleic acids cover Deoxyribonucleic Acid (DNA) and Ribonucleic Acid (RNA). They are tremendously important for life in general. DNA and RNA are long linear polymers and compose the bio-molecular family of nucleic acids. They consist of multiple nucleotide monomers. Each nucleotide itself is composed of a 5-carbon sugar attached to one or more phosphate groups and a nitrogenous base [3] (see Figure 1.1 and Figure 1.2). The polymer is called DNA if the sugar is deoxyribose. If the sugar is ribose, the polymer is called RNA [16]. The nucleo-bases connected to the sugars are Guanine (G), Adenine (A), Thymine (T), or Cytosine (C) for the case of DNA. Thymine is substituted for Uracil (U) in RNA. Cytosine, thymine, and uracil are pyrimidines as they contain the six-membered pyrimidine ring. Adenine and guanine are double-ringed purines containing a five-membered imidazole and a pyrimidine ring [3]. Each nucleotide is composed of a nitrogen-containing nucleobase as well as a monosaccharide sugar called deoxyribose and a phosphate group. The nucleotides are joined to one another in a chain by covalent bonds between the sugar of one nucleotide and the phosphate of the next, resulting in an alternating sugar-phosphate backbone (see Figure 1.1 for structure of bases). According to base pairing rules (A with T, C with G, and U with A), hydrogen bonds bind the nitrogenous bases of the two separate polynucleotide strands to make double-stranded DNA [3].

In particular, DNA is important as its sequence of the bases along the nucleic acid chain contains genetic information that is necessary for all known living organisms and viruses. The instructions are encoded as genes which are single units of genetic information packaged in a sequence of DNA. Specifically, the genes do not contain the information for the protein synthesis directly, but by using messenger-RNA they convey the genetic information to subsequently synthesize proteins [16]. The role of RNA is more complicated as it is involved in many processes from coding, decoding and regulation to the expression of genes. RNA and DNA are structurally different. In contrast to the double-stranded DNA, RNA can exist as a single-stranded molecule and a double-helix structure (see Figure 1.1 for comparison).

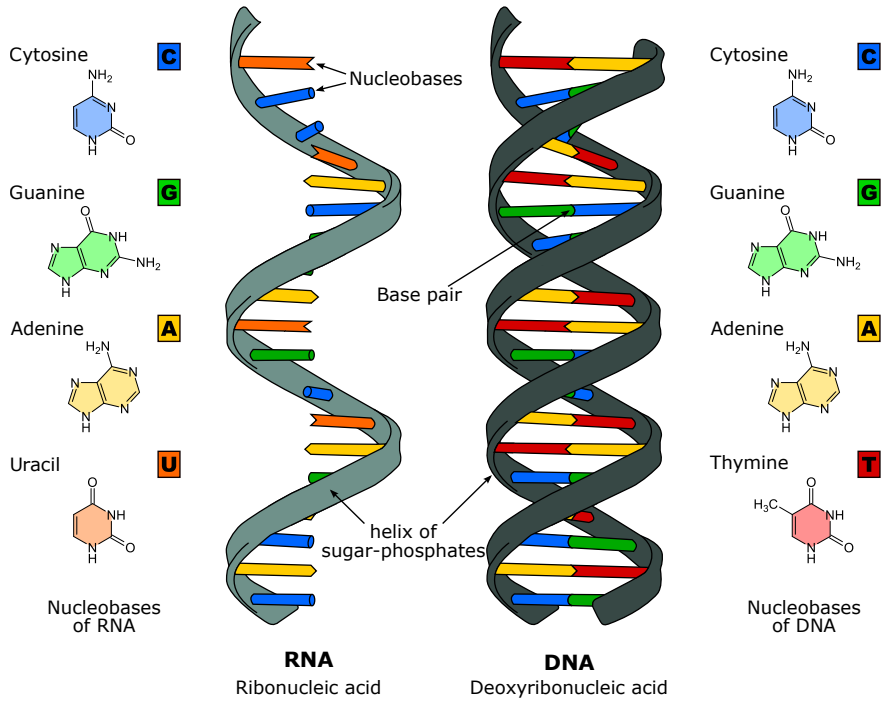


Figure 1.1: Difference between DNA and RNA. From [156].

Helical Structure of DNA

Friedrich Miescher first identified and isolated DNA in 1871. By choosing leukocytes as his source material, he aimed to investigate the proteins in these cells. To his surprise, he encountered a substance wit unexpected properties. This substance was DNA [34].

With the help of experimental data collected by Rosalind Franklin and Maurice Wilkins, James Watson and Francis Crick famously discovered the double-helical structure of DNA in 1953 [176]. Prior, it was assumed that the backbone of DNA faces inwards with the bases facing outwards to facilitate simple recognition processes. Watson, J. D.; Crick showed that this is not the case and that the structure of DNA can be depicted similarly as in Figure 1.1.

The DNA backbone is not symmetrical but has a specific direction due to chemical polarity. The polarity can be referred to by the 3′ end and the 5′ end [3]. The 3′ end is defined by a 3′ hydroxyl group and has a positive charge. The 5′ phosphate is negatively charged. This dipole moment gives the DNA backbone its direction (see Figure 1.2). These two chains are oriented in opposite direction to build the double-helical structure of the DNA. The two chains are held together by hydrogen bonds between the nucleotide bases opposing each other. These base pairs will be referred as Watson-Crick pairs (WC-pairs) in the following. On the larger scale the DNA is not only a double helix but it has a specific larger scale structure. Instead of being a straight double helix, DNA shows a coiled structure. It is tightly

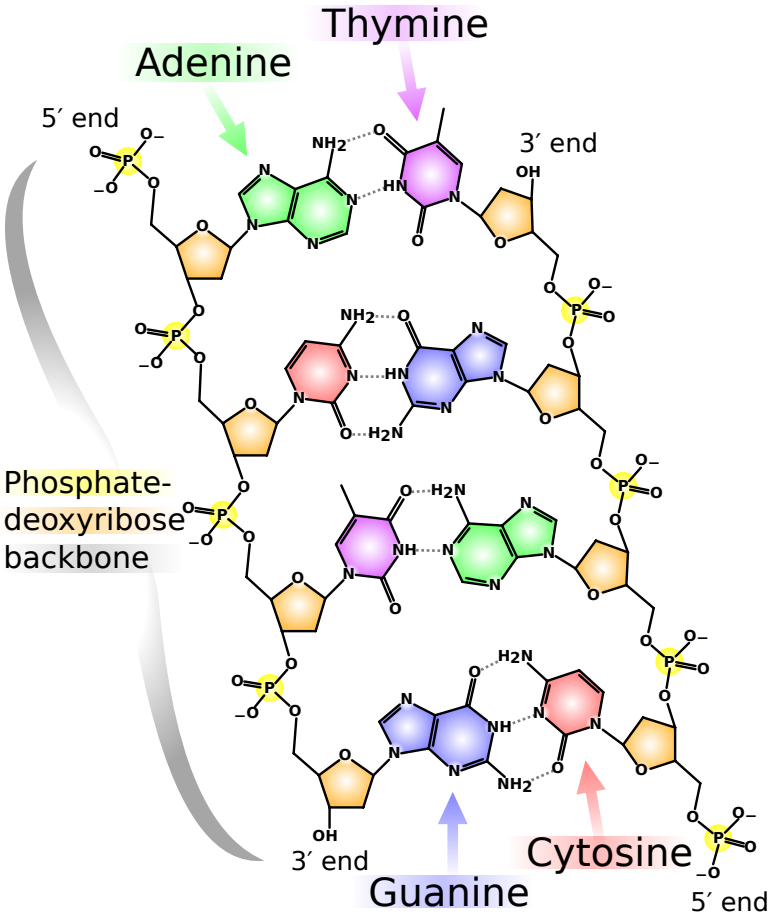


Figure 1.2: Structure of DNA and its polarity. From [111].

wrapped and super-coiled around eight so-called histone proteins [3]. Together they constitute the nucleosome core. The structure was first proposed by Kornberg and Lorch in 1974 (review paper: [91]). Later, the crystal structure was resolved by Luger et al. [109]. The Protein DataBase (PDB) structure PDB:1AOI is shown in Figure 1.3.

Double Stranded Structure of DNA

The double-stranded nature of DNA gives rise to many of its unique properties. Additionally to the Watson-Crick base pairing, other mechanisms such as *base stacking* make DNA resilient against damaging agents. *Base stacking* refers to the non-covalent bonds between the aromatic ring of DNA bases. Surprisingly, the DNA stability is mainly determined by the stacking interactions [181].

Nevertheless, DNA strands can still be separated. In particular, it is essential for all processes of DNA replication and some processes of DNA repair [3].

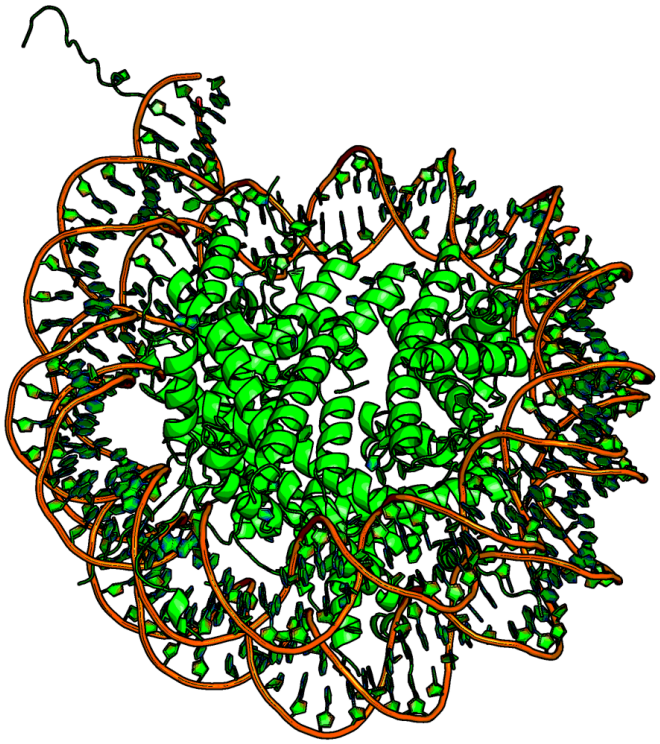


Figure 1.3: Crystal structure of the nucleosome core (PDB:1AOI). The green cartoon shows the H2A, H2B, H3, and H4 core histone [109]. Around 145 base pairs of the DNA are wound around the core histones.

Structural Features of DNA

The most commonly found conformation of DNA, B-DNA, is not the only possible conformation. DNA can also exist in at least two other states: A-form DNA (A-DNA) and Z-form DNA (Z-DNA). These forms differ in many aspects. Both, A-DNA and B-form DNA (B-DNA) are right-handed double helical structures.

The canonical form of DNA, B-DNA makes one helical turn approximately every 10.5 base pairs. The double-helix of B-DNA builds a cylinder of a diameter of 20 Å. The distance between the base pairs (rise) is 3.4 Å [176]. However, the exact parameters depend on the sequence of the given DNA. Another important aspect is the configuration of the deoxyribose sugar. This property is more specifically called the *sugar-pucker* [1]. *Endo* describes the configuration where the C2′ or C3′ are turned out of this plane into the direction of

1 The deoxyribose sugar is a 5-membered non-co-planar ring. It can therefore have multiple configurations, the so called pucker conformations.

O5′. The *exo* configuration describes a shift in the opposite direction. The sugar-pucker is in C2′ endo conformation in B-DNA [138, 166].

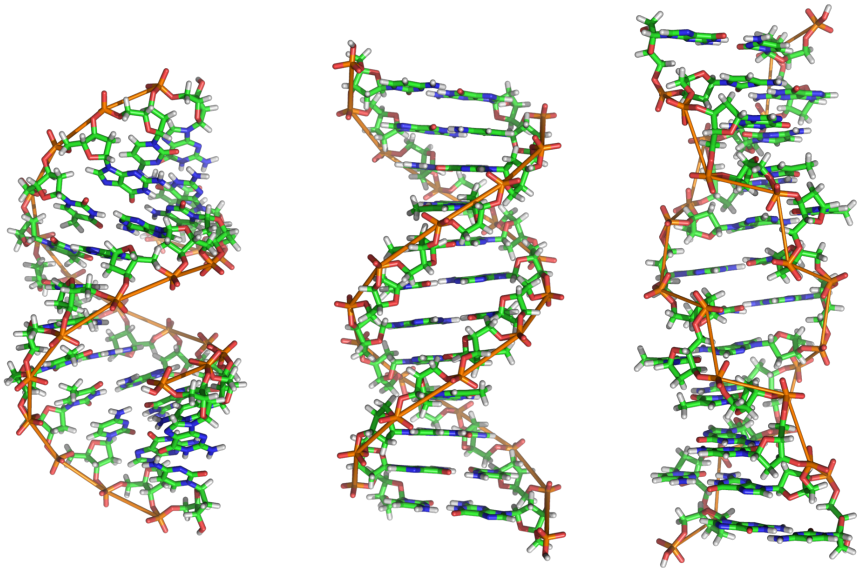


Figure 1.4: Comparison of A-DNA, B-DNA and Z-DNA (from left to right). From [183].

The second most important type of DNA is A-DNA. It can appear in environments of reduced water content or high salt content [138]. It has smaller grooves, 11 base pairs per turn and a tilt angle of the bases of 20°. In A-DNA, the sugar-pucker is in the C3′ endo configuration. Due to the base-pairs being shifted, A-DNA has a relatively central large hole of 9 Å viewed along the helical axis. The diameter is therefore larger with 23 Å. The rise is conversely smaller with 2.6 Å [138] (see Figure 1.4).

The Z type form of DNA differs significantly from the B type [172]. Z-DNA can be found in rare circumstances of high salt environments [136, 164] and was first discovered by Mitsui et al. [122]. In contrast to A- and B-DNA it has left-handed double helical structure and its backbone winds in a zig-zag pattern [67] giving this conformation its name (see Figure 1.3).

Backbone angles of DNA

As previously stated, the DNA backbone is limited in its structural degrees of freedom. Each type of DNA conformation is well defined by 6 torsion angles of the sugar phosphate backbone and by the angle describing the orientation of the base about the glycosidic bond [2].

2 A glycosidic bond is a type of covalent bond which joins a carbohydrate (sugar) molecule to another group. Glycosidic bonds can be formed with virtually any hydroxylated compound. The glycosidic bond is formed between the anomeric carbon of one monosaccharide and a hydroxyl group of another. The anomeric

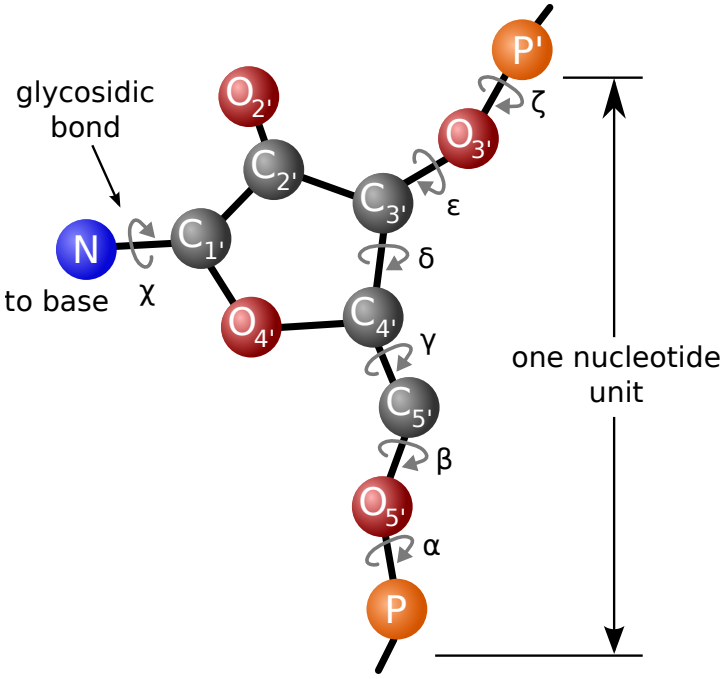


Figure 1.5: Dihedral angles of sugar-phosphate backbone of DNA. Adapted from [12].

These 7 angles are shown in Figure 1.5. B-DNA can obtain two possible conformations. The B_I and B_{II} regions define the two most probable conformations [83, 147]. A transition from the usual B_I state to the B_{II} state results in major changes of the overall DNA structure. It has been noted that changes in the base structure can lead to changes of the backbone angles, thus transmitting information from the bases to the phosphate groups [40]. This is important for an indirect readout and damage recognition upon protein binding[3]. The B_I and B_{II} states are usually defined by the difference of ϵ and zeta. For angles of $\epsilon - \zeta \approx -90°$ the backbone is in the B_I state, for angle-pairs of $\epsilon - \zeta > 0°$ it is in B_{II} [40].

Grooves

The double-stranded structure of DNA winds specifically to form grooves between the backbones winding around itself. The grooves are located between the strands. The wider groove is the major groove, whereas the smaller and narrower groove is the minor groove. Both groove can be sites for protein binding. For B-DNA, the major groove is roughly 11 Å wide and the minor groove is 6 Å wide. The

carbon is a stereo-center which in turn is an atom bearing groups such that an interchanging of any two groups leads to a molecules with the same sequence but different three-dimensional structure (stereo-isomer) [32, 115, 125].

3 Due to the double-helical structure the bases are not easily readout from the outside by repair proteins.

depth of the major groove is larger with roughly 4 Å in comparison to the minor groove with approximately 4 Å [158]. The larger width and depth of the major groove makes it the preferred binding site for protein binding. For one, transcription factors make contact with the side of the bases exposed in the major groove [131]. Other structural types of DNA such as A-DNA and Z-DNA are different in terms of groove parameters.

1.2 DNA DAMAGE

DNA can be damaged in multiple ways through many processes. A multitude of damaging agents can cause structural changes in DNA. These DNA damages are also called lesions. Depending on the type of cell and type of damage, lesions can occur 10^3 to 10^6 times per cell per day [17].

DNA damage and mutations should not be mixed up. While DNA damage can ultimately lead to mutations, DNA damage is basically a structural change in DNA that is not propagated itself whenever the DNA is replicated [17].

There has been the theory that DNA damage in non-replicating cells like brain cells and muscles is the main cause of aging in mammals [18, 69]. A possible mechanism might involve higher DNA damage triggering cellular signaling pathways, such as apoptosis[4] resulting in a faster depletion of stem cells which in turn contributes to accelerated aging [48]. Finally, unrepaired DNA lesions can lead to mutations which change the sequence of the DNA. They can then no longer be repaired as the base-pairing is valid. Mutations can also occur in the process of DNA replication.

It should also not be forgotten that even though "mutations or deficiencies in repair can have catastrophic consequences, causing a range of human diseases, mutations are nonetheless fundamental to life and evolution" [51].

Damage Types

Lesion Induced by Oxidization

One of the most common types of DNA damage is caused by the action of oxidizing agents mostly produced by cell metabolism. These lesion can ultimately result in mutations subsequently lung cancer [106]. Further, chemicals, Ultra-Violet (UV) and ionizing radiation can also result in reactive oxygen species drastically increasing the rate of oxidization in DNA [105].

Specifically, the oxidation effect can lead to many forms of of chemical modifications. These modifications include ring nitrogens

4 Apoptosis is the process of Programmed Cell Death (PCD).

and groups of nucleobases. 8-oxo-guanine is the most prominent example of such a modification and is one of the most abundant types of damages with one in 10^6 guanine bases being damaged in this form [185].

Lesion Induced by Irradiation

Photonic irradiation and other types of radiation can damage DNA by chemically altering its structure. Radioactive radiation such as gamma radiation leads to damage through the ionization process directly cleaving bonds in the molecular structure of DNA. The lower energy of UV light can still cause dramatic damage by exciting the DNA into a energetically higher state. It can relax by chemically changing its structure, i.e. building additional inter-atomic bonds. Lesion induced by ultra-violet light will be covered in more detail in Section 1.3.

Strand Breaks and Cross-links

X-ray irradiation can cause serious damage to DNA in form of single strand or double strand breaks. Single strand breaks can also occur by other types of damaging agents.

Radicals, alkylating agents and spontaneous reactions can all cause a single DNA strand to break [20]. Cisplatin[5] , Mitomycin C[6] and alkylating agents, which are used in chemotherapy, can cause DNA cross linking, causing either the same strand or opposite strands to form chemical bonds [72].

Missing Bases and Base Mismatches

Mismatched bases can occur in the process of replication and recombination. This process has been heavily studied for Escherichia coli [66, 77, 94]. Further, mismatches can occur as a result of other types of damages such as 8-oxoguanine, O^6-methyl-guanine, carcinogen adducts, UV photo-products and cisplatin adducts [77]. Missing bases also occur intentionally as a result of base excision repair.

1.3 DNA REPAIR

Although first indirect observations of DNA damage repair were done earlier, the DNA damage repair was first discovered in the 1940s by Friedberg [51].

Different types of DNA damage require specific repair mechanisms. For single base damages such as mismatches or oxidized bases, the

5 Cisplatin, cis-Diamminedichloroplatinum(II), has been widely used in chemotherapy for close to 30 years [31].

6 Especially guanine residues can be cross-linked by Mitomycin C [38].

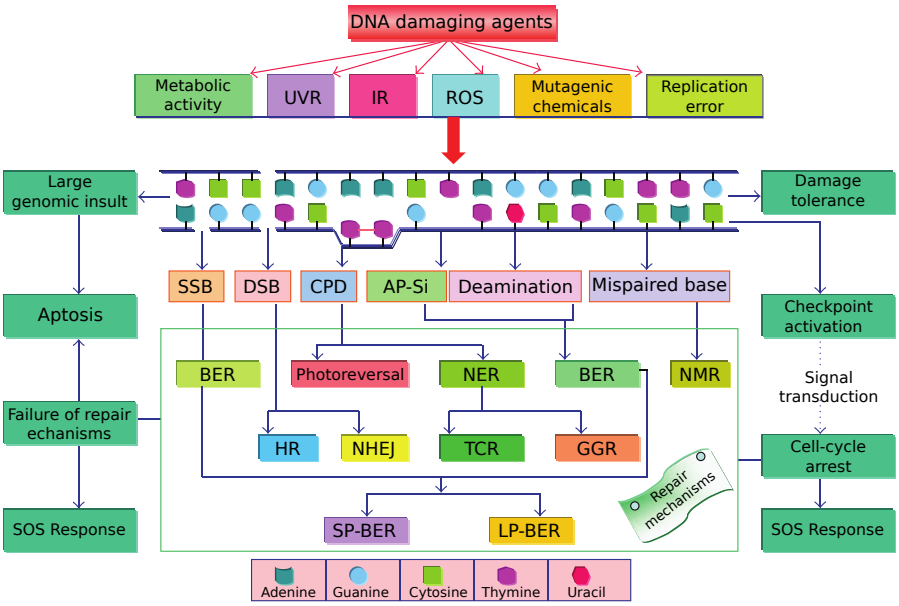
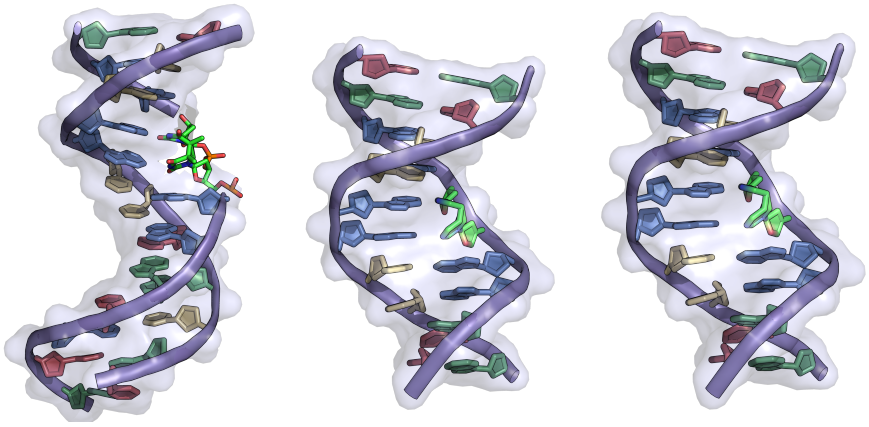


Figure 1.6: DNA damage and repair mechanisms. DNA lesions can be induced by various agents. The DNA can subsequently trigger repair mechanisms. If not repaired or in the case of very severe damage the lesion can lead to cell apoptosis or the SOS response. Damage can also be tolerated if not problematic for the genome. Multiple repair mechanism are shown for specific damage types [140]. The following abbreviations have been used: Ultra-Violet Radiation (UVR); Infra-red Radiation (IR); Reactive Oxygen Species (ROS); Single Strand Break (SSB); Double Strand Break (DSB); Abasic site (AP site); Base Excision Repair (BER); Nucleotide Excision Repair (NER); MisMatch Repair (MMR); Homologous Recombination (HR); Non-Homologous End Joining (NHEJ); Transcription Coupled Repair (TCR); Global Genome Repair (GGR); Short-Patch BER (SP-BER); Long-Patch BER (LP-BER); Save Our Soul response (SOS response); Programmed Cell Death (PCD). Figure is adapted from [140].

possible mechanisms are NER, BER, and MMR [51]. NER removes bulky helix-distorting lesions such as adducts of many xeno-biotics [7] to DNA bases [185]. MMR corrects errors made by DNA-polymerases during replication, excising canonical nucleotides incorporated into mismatches, as well as small insertion/deletion loops, from the daughter DNA strand [77]. NHEJ is a repair pathway for double-strand breaks in DNA. In contrast, homologous recombination repair need a homologous template, but can repair lesions that cannot be repaired by other repair mechanisms [124, 185] as well. Base excision repair deals with some of the most common DNA lesions such as oxidized bases, alkylation, deamination, base loss and single-strand breaks [185].

Repair mechanisms which can work without destruction of the damaged base and a subsequent DNA re-synthesis are commonly

7 Xeno-biotics are foreign chemical substances in biologic systems.



(a) PDB:1TTD [113]. (b) PDB:1N4E [133]. (c) The dewar of the 6-4PP, PDB:1QKG [96].

Figure 1.7: Crystal and NMR structures for different types of DNA damage. Figure 1.7a and Figure 1.7b are DNA containing CPD photo-lesions.

referred to as direct reversal repair processes [185]. In this thesis, the focus will be on the recognition and repair of cyclobutane pyrimidine dimers with the help of visible light[8]. Figure 1.6 shows an overview of typical damages and their repair pathways.

Cyclobutane Pyrimidine Dimer (CPD) and 6-4 Photo-products

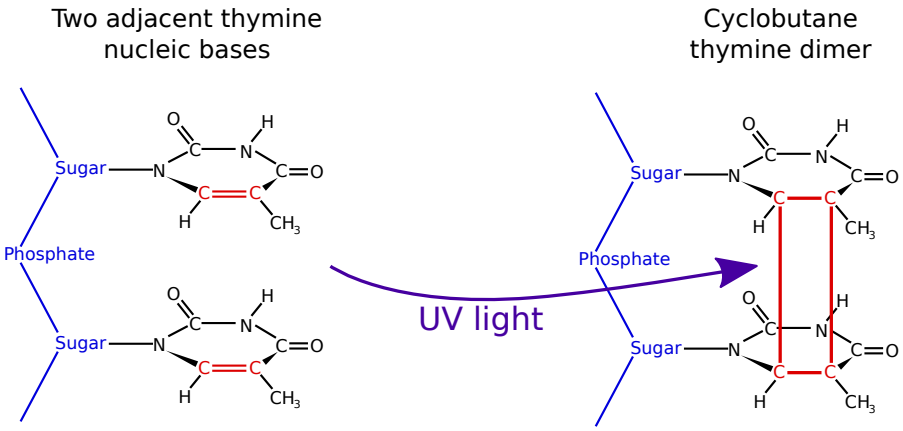


Figure 1.8: CPD damage formation. Not true to scale as the two additionally formed bonds are longer in comparison two the base rings.

The two major types of DNA lesions resulting from UV radiation are CPD and 6-4PPs [149]. CPD damage is more common with a fraction of 75 % in comparison to 6-4PP lesions with 25% of the total UV-caused DNA lesions. Both types of damage distort the DNA

8 Cyclobutane is a cycloalkane and an organic compound with the chemical formula C_4H_8 [19].

helix [149]. In comparison, 6-4 photo-products change and distort the DNA structure more than cyclobutane pyrimidine dimers. For 6-4PP lesions a strong bending of 44° is observed [14, 85] This results in a faster and more efficient repair process in comparison to the repair of CPD lesions. 6-4PP lesions are removed approximately five to ten times faster from DNA in vitro [14, 121] than CPD lesions.

CPD lesions corresponds to two additional bonds between the respective C5 and C6 atoms of the adjacent thymine bases [140], shown in Figure 1.8. The CPD lesion can occur in multiple different diastereoisomers (see Figure 1.9), i.e. configurations. The main adduct or dimer are cis-syn compounds. Too much smaller extent trans-syn lesions are formed, mostly by irradiation of single-stranded DNA [25]. The other forms of adducts occur in even more rare cases [140].

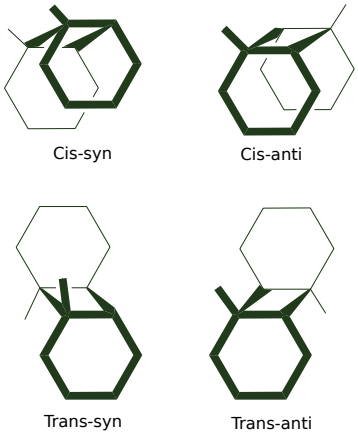


Figure 1.9: The dimerization of two thymine to CPD can occur in four ways. The cis-syn species is the most common dimer and will be the focus from heron. Adapted from [140].

1.4 CPD DAMAGE REPAIR

It has been shown that the CPD lesion, also termed thymine dimer damage, is, if not repaired, highly cytotoxic, mutagenic, and carcinogenic [140, 146]. Thus, an efficient repair process is tremendously important. CPD lesion can be repaired in multiple ways. The appropriate repair mechanism depends heavily on the specific organism and circumstance. The different domains of life, the archaea, bacteria, and eukaryote domain, have different and overlapping mechanisms of CPD damage repair. In general, CPD damage repair can be divided into light-dependent and light-independent mechanisms.

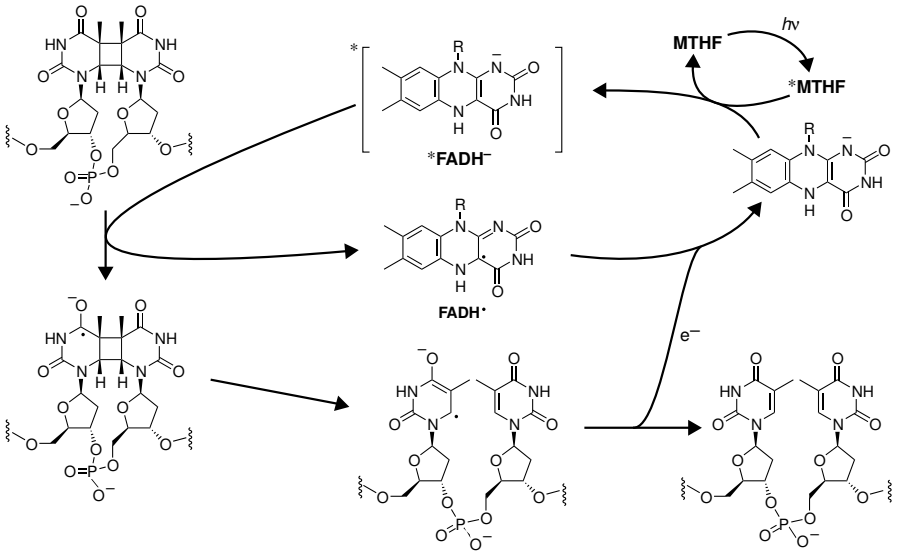


Figure 1.10: CPD repair mechanism by direct photo-reversal using photolyase. After the absorption of light by the methenyltetrahydrofolate (MTHF) cofactor, the energy is transferred via Fö,rster dipole-to-dipole transfer to the reduced and deprotonated Flavin Adenine Dinucleotide (FAD). The electron is then further transferred to the dimer initiating the photo-reversal [26]. Adapted from [26].

Light-dependent Repair

In bacteria CPD lesions and the related 6-4 pyrimidine-pyrimidine photo-lesion can be reversed by exposure of the bacteria to blue light. Using a light-induced reaction, these enzymes split the cyclobutane ring and restore the intact bases [21, 26, 35, 41, 127, 140, 146]. This process is referred to as direct photo-reversal [59]. It has been shown that this process is efficient in plants [22, 132, 163].

Photolyase is a repair protein which uses photo-reversal. It contains one FAD cofactor and a second co-enzyme, either a Methenyl Tetra Hydro Folate (MTHF) (type-I) or a 8-hydroxy-5-deazariboflavin (type-II) [26]. The MTHF cofactor is bound very closely to the surface of the protein [26]. The MTHF cofactor captures and absorbs light, transfers the excitation energy via Förster dipole-to-dipole transfer to the reduced and deprotonated FAD. Subsequent electron transfer to the dimer initiates the last step of photo-reversal [26] (see Figure 1.10).

Photolyases have a high sequence homology with cryptochromes, proteins which are very important to many aspects of life from bacteria to humans. One of their functions is synchronising the cardian rhythm with external stimuli from solar day cycles. Both proteins are thought to have a common origin in evolution [146].

Light-independent Repair

Although cryptochromes can be found in most eukaryotes such as mammals, this is not the case for photolyases. Importantly, photolyase and (6-4) photolyases are not present in placental mammals, including humans. The (6-4) photolyase has up to this point not been found in prokaryotes [70, 104, 146].

As suggested by [59] another method to repair CPD lesion can be used in many organisms which leverages so-called dark repair pathways, including NER. A concrete example is the NER repair process in yeast Saccharomyces cerevisiae [14, 52]. For plants, CPD lesions and 6-4 photo-products are often repaired by NER without the involvement of direct photo-reversal [22, 59, 167]. In mammalian cells the CPD lesion is repaired by the NER pathway with the Xeroderma Pigmentosum Complementation group Human Rad23 Homolog complex (XPC-hHR23B)-complex [1, 14, 64, 93, 99, 126, 185] forming the primary damage recognition protein, which initiates the NER pathway after damage binding [46, 93]. This in-vitro NER process in human cells is a complicated multi step procedure involving at least six core enzymes including the XPC-hHR23B complex, the 6-9 subunit Transcription factor II Human (TFIIH)[9] complex and two endonucleases, Xeroderma Pigmentosum, complementation group G (XPG)[10] and the Excision repair cross-complementation Group - Xeroderma pigmentosum, complementation Group F (ERCC1-XPF) complex[11] [14]. Several other DNA repair enzymes have been described that can recognize and repair photo-damaged DNA independent of light [26, 46, 50, 51, 93, 127, 146, 159, 160].

1.5 RECOGNITION

The focus of this thesis is to explain the recognition process of CPD damage repair by photolyase within Escherichia coli (E. coli). A major question remains: At which point in the repair process is the lesion recognized or is it recognized at all? One straightforward repair process flips out each base prior to attempting to repair it regardless of whether it is damaged or not. These repair processes would not depend on the distinction between damaged and undamaged DNA nucleobases.

Most repair processes such as BER rely heavily on recognition and selectivity. Because processes which happen later in the repair are not damage specific, selectivity is of tremendous importance.

9 TFIIH is responsible for the melting of the DNA and the formation of the open complex [47, 86].

10 Single-stranded structure-specific DNA endonuclease involved in DNA excision repair.

11 The ERCC1-XPF complex is essential for the repair of DNA damage in humans. It is a structure-specific endonuclease and uses nucleotide excision repair [116].

Without selectivity, BER would cut out every base regardless whether it is damaged or not. In BER, it is believed that the glycosylase step recognizes the damage and is thus responsible for selectivity [130]. During the course of this thesis, it was shown that a similar mechanism is used for CPD repair by photolyase from E. coli.

2 THEORY

2.1 INTRODUCTION TO MOLECULAR DYNAMICS - MD

Molecular Dynamics (MD) simulations are a type of computer simulation that calculate the physical movements of atoms and molecules. In standard MD simulation the movements and positions of the atoms are determined by solving the classical Newton's equations of motions. The resulting sets of atomic positions over multiple simulation steps yield the trajectories of these atoms and in turn of the molecules. The interactions between the atoms are modeled by simple potential functions whose parameters are given by so called *force fields* (chemistry)[1]. The method of MD was independently developed by Alder and Wainwright[5] and Rahman[139].

MD can be applied if analytical solutions are not available and experiments cannot yield the necessary information. Almost all simulated systems cannot be solved analytically. Even the solution of the simple but famous three-body system needs some use of numerical methods. Experiments on the other side have shortcomings in the measurement of the structure and the dynamics of molecules at the same time. Most experimental methods have either high spatial or high time resolution. For example, X-ray crystallography allows to determine the structure of a bio-molecule with very high spatial resolution. As this is a measurement over a long period of time, usually hours, the time resolution is very poor. Some other spectroscopic method such as Förster Resonance Energy Transfer or Fluorescence Resonance Energy Transfer (FRET) can measure distances of a few light-sensitive molecules - fluorophores - with high time accuracy in the order of 1×10^{-3} s–1×10^{-9} s. The method is however limited to a small number of fluorophores. A large number of fluorophores would alter the structure and dynamics of the systems drastically[2].

In MD, the resolution in time is only limited by the size of the time-step used. In the final implementation the time-step is set as high as possible without altering the characteristics of the systems to allow the study of motions on longer time-scales as every calculation for a time-step needs roughly the same amount of computational

1 force fields describe in our context the term used in chemistry not to be confused with force fields (physics) in classical physics. From here-on, only the definition of chemistry will be used.

2 The smallest fluorophores are in the range of 20 atoms and therefore still alter the dynamics of most small protein and ligands.

resources. A usual value of this time-step is 1 fs. This can capture the fastest motions, the movements of hydrogen atoms, accurately. The resolution in space is limited theoretically by the accuracy in saving these coordinates. In most simulation suites (see Appendix A), this is done in single precision. In reality, the precision here does not matter, as the accuracy of describing the forces present in the system is more limiting.

Typical time-scales which are simulated range from few nanoseconds for large systems to multiple microseconds for small systems. In this time-scale bio-molecules can show local motions, collective local motions and small subunits and domains can fold. The simulation effort and the compute-time needed range from seconds to weeks and from one Central Processing Unit (CPU) core to a complete cluster of machines.

Other types of simulations which are accurate to the level of Quantum Dynamics are only possible for short periods of simulated time. However, mixed approaches like Quantum Mechanics/Molecular Dynamics hybrid (QM/MD) where a quantum mechanical region is mixed with a MD simulation of the remaining part of the system, have proven to be successful for specific problems where chemical processes can occur.

2.2 HISTORY AND APPLICATION

After first successes of computer simulations in general and Monte Carlo simulations in particular, a numeric calculation of the rather theoretical example of an an-harmonic, one-dimensional crystal by Fermi, Pasta, and Ulam started the development of Molecular Dynamics simulations [43, 49]. The first proper Molecular Dynamics (MD) simulation was reported in 1956 by Alder and Wainwright [5] while the method was independently developed by Rahman in 1964 [139]. In 1960 Gibson et al. simulated a more realistic system of radiation damage in crystalline copper [58].

During the following years, MD simulations have been broadly applied to material science. One can argue that the underlying algorithms of MD have hardly changed since its beginnings in the 1950s [49].

With the vast improvement of computers in the following years, the method of MD simulation was applied to a vast range of problems. Specially noteworthy was the application of MD to the study of proteins by Levitt in 1976 [101] and by Warshel in 1976 [175] after envisioning the application of MD to protein folding in 1975 [102]. The simulation of the first protein in water was done in 1983 for a short simulation time of 2×10^{-11} s [103]. In 1997 the first peptide was simulated and folded for the much longer time-span of 1×10^{-7} s. A good overview about the historical account of MD and

its application to bio-molecules is given by Karplus and Mccammon [84].

Martin Karplus (Harvard), Michael Levitt (Stanford), and Arieh Warshel (USC) were awarded with the Nobel Prize in Chemistry in 2013 "for the development of multi-scale models for complex chemical systems." [56].

2.3 THE IDEA OF MOLECULAR DYNAMICS

Molecular Dynamics simulations are very similar in their execution to the approach in experiments. First, the system (in experiments: the sample) is prepared. Then, the simulation is run until the systems properties do not change dramatically over time. The system is therefore in the equilibrated state. In experiments this requirement means sufficient statistics. Now, the averages of observables (experiments: measurements) can be taken [49].

A simple algorithm for MD would look like this:

1. The system is initialized. Thus, all atoms are given positions and importantly starting velocities according to the overall system temperature.
2. The forces acting on each of the particles are calculated.
3. By integrating Newton's equations of motion, new positions and velocities are calculated. This step is repeated until the desired simulation length is reached.
4. Afterwards, averages of observables can be computed for purposes of analysis.

Each of these steps, and in particular step 3 and 4 are tremendously more complicated. The important details will be explained in the following.

2.4 BIO-MOLECULAR INTERACTIONS AND POTENTIALS

At the scale of our interest, only electro-magnetic interactions and the quantum mechanical exchange interactions are of importance. However, these interactions can be further categorized to simplify the numerical model and the understanding of the system.

Exchange interactions are the interactions of identical particles. For our purposes, only the Fermi repulsion for fermions plays a major role. It gives rise to strong repulsion effects between atoms. Due to the quantum mechanical nature, atoms can share electrons and reduce their collective energy. This interactions lead to covalent bonds. Together with electrostatic interactions between ions, these two terms are represented by the model of Van der Waals's force.

Van der Waals's interactions are represented together with the Fermi repulsion by Lennard Jones potentials. This approximation has proven to be reasonably accurate and is widely used in computer simulations [49]. The second type of basic interactions are electrostatic interactions which are of comparably long range nature.

In the model of MD, the involved interactions are modeled slightly differently. The model divides between three types of interactions. Inter-molecular interactions act between molecules. Such interactions are Lennard-Jones type interactions and Coulomb interactions. These kinds of interactions happen also intra-molecularly. Additional intra-molecular interactions due to chemical bonds such as covalent or ionic bonds cannot be broken in the framework of MD simulations. The third type of interaction is between molecules and the solution. This special type of interaction in particular will be explained in Section 2.9.

In Amber [28] the forces are described by the terms

$$V(\mathbf{r}) = \sum_{\text{bonds}} 1/2 k_{bi} \left(b_i - b_{i,0}\right)^2 \tag{2.1a}$$

$$+ \sum_{\text{angles}} 1/2 k_{\theta i} \left(\theta_i - \theta_{i,0}\right)^2 \tag{2.1b}$$

$$+ \sum_{\text{torsions}} \sum_{n=1..N i} k\tau_{ni} \left(1 + \cos\left[n_i \tau_i - \delta_i\right]\right) \tag{2.1c}$$

$$+ \sum_{\text{nb pairs}} \left[4\epsilon_{ij} \left[\left(\sigma_{ij}/r_{ij}\right)^{12} - \left(\sigma_{ij}/r_{ij}\right)^{6}\right] + q_i q_j/\left(4\pi\epsilon_0 r_{ij}\right)\right] . \tag{2.1d}$$

The first term describes the harmonic approximation of chemical bonding potentials (see Equation 2.1a). As most higher order quantum mechanical wave functions are not radially symmetric, chemical bonds in molecules are dependent on angles and dihedral angles of triples and quadruples of atoms. These terms are described in Equation 2.1b and Equation 2.1c. For non-bonded (written as nb in the formula above) atom interactions, the interactions are described by Lennard-Jones and electrostatic potentials (see Equation 2.1d). Other MD suites use interactions which are very similar.

2.5 FORCE FIELDS

The parameters of Equation 2.1 are different for each type of atom. Interaction parameters are not constant for specific atoms, but are further specialized for specific atoms in different contexts. As such, a Carbon atom would have different parameters depending on its position in the backbone of the protein. These parameters are generally bundled together as *force fields*.

The parameters of the force fields are usually determined empirically. As most sets of parameters have been used extensively in chemistry, a vast amount of experimental data is available for parameters such as atomic mass or Van der Waals's radii. A method for less-studied molecules includes the use of quantum mechanical calculations. Together with optimizations on parameters such as enthalpy of vaporization and sublimation, a force field can be developed.

In high-dielectric media such as water polarization effects play a crucial role. Recently, polarizable force fields have made improvements in this field[3]. These force fields are still under heavy development and have not been used in this thesis.

To reduce the quantum mechanical description of the potentials to the much simpler classical model mentioned above, multiple approximations are used. The first one is the Born-Oppenheimer approximation which states that the dynamics of electrons is so fast that they can be considered to react instantaneously to the motion of their nuclei. As a consequence, they may be treated separately. The second approximation treats the nuclei which are much heavier than electrons, as point particles that follow classical Newtonian dynamics. In classical molecular dynamics the effect of the electrons is approximated as a single potential energy surface, usually representing the ground state.

2.6 EQUATION OF MOTION

As stated before, Newton's equation of motion has to be solved in order to calculate the updated velocities and positions after one step of simulation. Newton's equation of motion is given by

$$\mathbf{F}_i = m_i \mathbf{a_i} \,,$$

where the force is given by the negative gradient of the potential.

$$\mathbf{F}_i = -\boldsymbol{\nabla} V_i \,,$$

This can be simply combined to

$$m_i \mathbf{a_i} = -\boldsymbol{\nabla} V_i \,,$$

and as $\frac{d^2\mathbf{x}}{dt^2} = \mathbf{a}$ the equation of motion is given as

$$m_i \frac{d^2\mathbf{x}_i}{dt^2} = -\boldsymbol{\nabla} V_i \,.$$

3 Details about these developments can be obtained from the review paper of Antila and Salonen [10]. Another good overview is given by Mackerell [110].

As the equation of motion is calculated in a step-wise manner, the integration to yield the new velocities and positions of all atoms is simple. A correct time-reversible solution is symmetric in its expansion around the current time t. A Taylor expansion of the position vectors around the time $t + \Delta t$ yields

$$\begin{aligned}\mathbf{x}_i(t+\Delta t) &= \mathbf{x}_i(t) + \frac{d\mathbf{x}_i(t)}{dt}\Delta t + \frac{d^2\mathbf{x}_i(t)}{dt^2}\frac{\Delta t^2}{2} \\ &\quad + \frac{d^3\mathbf{x}_i(t)}{dt^3}\frac{\Delta t^3}{6} + \mathcal{O}(\Delta t^4) \\ &= \mathbf{x}_i(t) + \mathbf{v}_i(t)\Delta t + \mathbf{a}_i(t)\frac{\Delta t^2}{2} \\ &\quad + \frac{d^3\mathbf{x}_i(t)}{dt^3}\frac{\Delta t^3}{6} + \mathcal{O}(\Delta t^4)\,.\end{aligned}$$

This can be done similarly for the time-step before the current time as

$$\begin{aligned}\mathbf{x}_i(t-\Delta t) &= \mathbf{x}_i(t) - \mathbf{v}_i(t)\Delta t + \mathbf{a}_i(t)\frac{\Delta t^2}{2} \\ &\quad - \frac{d^3\mathbf{x}_i(t)}{dt^3}\frac{\Delta t^3}{6} + \mathcal{O}(\Delta t^4)\,.\end{aligned}$$

By adding these two equations and solving for the $\mathbf{x}_i(t-\Delta t)$ term, the following time-reversible equation is obtained

$$\mathbf{x}_i(t+\Delta t) = 2\mathbf{x}_i(t) - \mathbf{x}_i(t-\Delta t) + \mathbf{a}_i(t)\Delta t^2 + \mathcal{O}(\Delta t^4)\,.$$

This is also known as the Verlet algorithm which was first derived in 1791 by Delambre, but famously rediscovered by Verlet in 1967 [171].

Another solution using directly the velocities gives

$$\mathbf{x}_i(t+\Delta t) = \mathbf{x}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{1}{2}\mathbf{a}_i(t)\Delta t^2\,,$$

and

$$\mathbf{v}_i(t+\Delta t) = \mathbf{v}_i(t) + \frac{\mathbf{a}_i(t+\Delta t) + \mathbf{a}_i(t)}{2}\Delta t\,,$$

which is known as the velocity-Verlet algorithm [178] and has the advantage of not needing to save the past positions of all atoms $\mathbf{x}_i(t-\Delta t)$. Thus the simplest approach of solving the equations of motions is as follows and commonly used in Molecular Dynamics:

1. Calculate the forces $\mathbf{F}_i(t)$ on each atom in the system (due to all interactions with other atoms) given the current coordinates $\mathbf{x}_i(t)$ and velocities $\mathbf{v}_i(t)$.

2. Update the positions of the atoms according to $\mathbf{x}_i \leftarrow \mathbf{x}_i + \mathbf{v}_i\Delta t + \frac{\mathbf{F}_i}{2m_i}\Delta t^2$

3. As the forces can also depend on the velocity, the second equation cannot be solved perfectly. Approximately, it is solved by calculating the second equation partially and neglecting the $\mathbf{a}_i(t + \Delta t)$ term. Thus the velocities are updated as $\mathbf{v}_i \leftarrow \mathbf{v}_i + \frac{\mathbf{F}_i}{2m}\Delta t$
4. Calculate the new forces $\mathbf{F}_i(t + \Delta t)$ using the current positions $\mathbf{x}_i(t + \Delta t)$.
5. As the forces are now overwritten, updating the velocities as $\mathbf{v}_i \leftarrow \mathbf{v}_i + \frac{\mathbf{F}_i}{2m}\Delta t$ delivers the correct term according to the second equation.

This procedure is repeated for each simulation step [49].

2.6.1 *Shake Algorithm*

As atomic bonds are not represented in terms of bond-potentials but rather by hard constraints, these constraints are treated as Lagrange multipliers in the Hamiltonian or Lagrangian formalism. By neglecting some intra-molecular motions, the tiny fluctuations of hydrogen atoms, the time step can be drastically increased. The Lagrange multipliers and therefore the constraints are violated as only the equations of motions solved approximately [6]. The Verlet algorithm solves this problem by satisfying the conditions exactly at the end of each simulation time step[171].

The specific implementation of the Verlet algorithm in most MD simulation software systems is the SHAKE algorithm [143]. The constraint of the before-mentioned Lagrange multipliers are implemented with the Gauss-Seidel method. This method is iterative and similar to the well-known Jacobi method. For velocities, the Algorithm for Rigid Water Models (RATTLE) algorithm can be used. On computers with fixed precision, the RATTLE algorithm is more accurate than SHAKE. It is also easier to modify RATTLE for constant temperature or pressure simulations [7]. As later explained, constant temperature/pressure MD is highly desirable as those most accurately describe the modeled systems.

2.7 PERIODIC BOUNDARY CONDITIONS

Periodic Boundary Conditions (PCB) are boundary conditions which are used to approximate a very large or even infinite system by a small part of it. This is done by replicating this small part, the *unit-cell*, infinitely. In the implementation, however, only one unit-cell has to be simulated. A particle escaping on one side of the unit-cell reappears on the opposite side of the same unit-cell. Same holds true for all interactions.

The simplest example of periodic boundary conditions in three-dimensional space are cubic periodic boundary conditions

$$\mathbf{x} = \mathbf{x} + \mathbf{a}_p \,, \quad p \in 1,2,3 \,,$$

and

$$\mathbf{a}_1 = (a,0,0) \,, \quad \mathbf{a}_2 = (0,a,0) \,, \quad \mathbf{a}_3 = (0,0,a) \,.$$

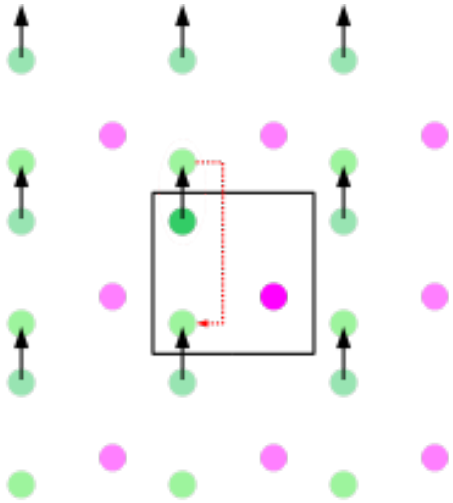


Figure 2.1: Periodic boundary conditions in two dimensions [63].

In general periodic boundary conditions have to be space-filling. Additionally, in MD, only one type of unit cell can be used. The combination of multiple unit-cells would make matters too complicated. Thus, polyhedral structures like the *Weaire-Phelan* structure with two types of cells [55] cannot be implemented.

One common polyhedral structure for periodic boundary conditions is the truncated octahedron. It has the advantage of resembling a sphere very closely. The solvent can move and rotate freely in the unit cell as its distance to its copy in the periodic unit cell has to be above a critical distance in order to avoid problems with long range interactions such as electrostatics. In other applications such as the simulation of a membrane, a cubic or cuboid is of better use[4].

2.8 PARTICLE MESH EWALD

In the method of explicit solvent our systems are treated periodically, the long range electrostatic interactions will then be calculated by

4 Setups which are highly non-spherical such as long proteins can also benefit from cuboid unit cells. In these cases some kind of restraint has to be employed to avoid the protein interacting with itself

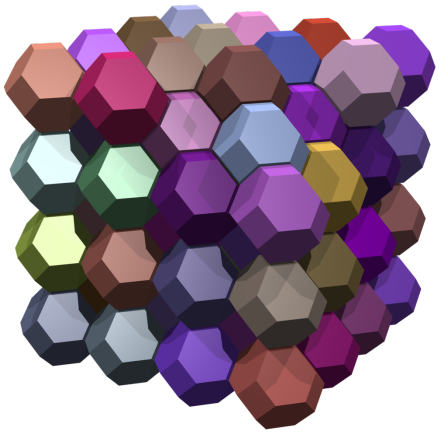


Figure 2.2: Truncated octahedron [8].

Ewald-summation, named after Paul Peter Ewald. If the system had been treated in a non-periodic manner, cut-off artifacts would be introduced [90, 100, 144, 145]. Thus, the standard Coulomb's law has to be modified in order to avoid such artifacts for finite distance calculations.

The standard Ewald algorithm is computationally quite expensive as its computational effort grows with $N^{3/2}$, N being the number of charges [100].

The Particle Mesh Ewald (PME) method of Darden, York, and Pedersen reduced this to a computational effort of $O(N\log(N))$ [37]. In particular, the Ewald sum can be modified such that the direct space sums can be computed in the order of N. Then, the remaining sum in reciprocal space can be approximated by the particle-mesh interpolation approach of Memon, Hockney, and Mitra [118]. The energies are now convolutions which can be quickly computed using Fast Fourier Transforms (FFTs) [37]. An algorithm of this method scales with $N \cdot \log(N)$. The introduction of this method helped to scale MD simulations from few thousands into millions of atoms.

2.9 EXPLICIT VERSUS IMPLICIT SOLVENT MODELS

For the explicit simulation of water the TIP3P [79] water model is commonly used. It is a simple, inflexible three sided model. The TIP3P model is used throughout the course of this thesis.

Instead of simulating every solvent explicitly, the Generalized Born (GB) model can also be used to implicitly simulate the solvent. This model has proven to be accurate for non-polarizable force field such as the ff99SB and ff03 [28]. Here, the total solvation free energy of a molecule, ΔG_{solv} is assumed to be decomposable into an electrostatic and a non-electrostatic term.

$$\Delta G_{solv} = \Delta G_{el} + \Delta G_{no\text{-}el}\,.$$

Ensemble	Fixed quantities	Thermodynamic potential
Micro-canonical	N, V, U	S maximal
Canonical	N, V, T	F minimal
Isothermal-isobaric	N, P, T	G minimal
Grand-canonical	V, T, μ	Ω minimal

Table 2.1: Statistical ensembles and their associated thermodynamic potentials[4].

$G_{\text{no-el}}$ is the solvation free energy of a molecule from which all charges have been removed, and the free energy $G_{\text{no-el}}$ can be calculated by first removing all charges in the vacuum, and then adding them back in the presence of a continuum solvent environment [28]. An analytic GB [28] method is used in Assisted Model Building with Energy Refinement (AMBER) to simplify the numerical computation of the Poisson-Boltzmann equation with reasonable accuracy. As implicit models have not been used in the course of this thesis, further details are omitted.

2.10 THERMODYNAMICS AND STATISTICAL DYNAMICS

Computer Simulations and MD simulations in particular can record coordinates and velocities of atoms with great precision. However, as fine-grained measurements of all atoms are not possible, this cannot be compared to experiments. Experiments would measure averaged properties such as temperature and pressure. To calculate how the microscopic properties are connected with macroscopic properties averaged over large number of particles, methods from statistical mechanics are used.

An ensemble in statistical mechanics is defined as the set of all possible states a system can adopt. Depending on the fixed external variables, different thermodynamic ensembles are defined. A system with fixed internal energy U, fixed volume V and fixed number of particles N is called the micro-canonical ensemble. If the temperate T, volume V and number of particles N are fixed, the canonical ensemble is given. The isobaric-isothermal ensemble has fixed N, P,and T and is often used in MD as well. Many experiments have similar conditions making this ensemble a realistic description of many systems. For fixed T, V and chemical potential μ, the grand canonical ensemble describes the system.

Each ensemble can be described in equilibrium by an associated thermodynamic potential similar to the Hamiltonian in Quantum Mechanics. The ensembles, their fixed macroscopic quantities, and their thermodynamic potentials are listed in Table 2.1 [4].

For the micro-canonical ensemble the associated thermodynamic potential is the entropy S which is maximized under those conditions. In most simulations the canonical ensemble is used as it is most easily modeled. It is also in realistic agreement with many experiments under fixed number of particles, such as a cell. Here the free energy F is minimized. As such it is in interesting property to study.

Ergodicity

Even if only a single system is observed, the theory of statistical ensembles can be used to make predictions about observables. This is possible due to the ergodic hypothesis which states that averages over the ensemble are equivalent to the averages over a very long time. As such, averages can either be computed by time-averaging (MD approach) or by ensemble averaging (Monte Carlo (MC) approach)[49]. The ergodic hypotheses holds if the measured time is of a much larger scale than the equilibration time of the system. A system which obeys the hypothesis is ergodic. In practice there are many systems which are not ergodic such as glasses and meta-stable phases [49].

2.11 LIMITATIONS OF MOLECULAR DYNAMICS

Due to various limitations MD simulations cannot be completely accurate. In addition to the approximations of the algorithm itself, the model of the system introduces many artifacts Such problems occur due to the choices of how the system is modeled: 1) Which interactions are included, 2) how those interactions are described by force fields, 3) how the degrees of freedom are sampled, 4) how the boundaries of the system are modeled [169], 5) how long the system needs to be simulated, and 6) how system is compared to experiments.

These questions lead to the distinction of the limitations into four categories [169]: The force field problem, the search problem, the ensemble problem and the experimental problem.

The Force Field Problem

Due to the summation of errors, it is inherently difficult to describe a large system with force fields at high accuracy. Additionally, entropic effects are hard to account for in force fields. Further, the vast amount of parameters make force field generation a tremendously difficult task [169].

A macroscopically large system cannot be calculated computationally. Even simulating one million number of atoms, the size of the system is only a small fraction in regard to Avogadro's number [169].

Thus, simplifications have to be used. The simulation in vacuum is fast but does not account for surface effects and the dielectric properties of the solution. Small solution droplets around the simulated bio molecule lead to problems such as surface effects at the boundaries of the droplet and evaporation of the solution itself. Periodic systems have proven to offer the best solution for explicit simulations but introduce artificial periodicity [169]. This problem can be reduced with appropriate system sizes. Other solutions, such as implicit models have been described here but introduce other problems.

The Search Problem

The very large number of degrees of freedom of usual MD systems is problematic by itself. Even small systems have number of degrees of freedom in the range of 1×10^4–1×10^6. Thus, the energy surface of these systems is very rough. Even searching for the global energy minimum becomes problematic [169]. To search the configuration space for important contributions, a wide range of methods is available. Some of these methods will be covered later in Section 2.12.

The Ensemble Problem

Due to statistical considerations, the global potential energetic minimum is not sufficient for almost all considerations. Entropic effects are inherently important. Therefore the configuration space has to be sampled in accordance with its contributions and the associated Boltzmann factors.

The Experimental Problem

Comparison of MD simulation data with experiments presents itself as another problem. Most experimental observables are averages over many particles or a long period. Additionally, a system with a high number of degrees of freedom is measured using only few observables. Thus, different effects may contribute to the same outcome in these observables. Last but not least, experiments are often not sufficiently accurate making comparisons to MD simulations very tricky [169].

2.12 FREE ENERGY AND ENHANCED SAMPLING METHODS

As the free energy is a observable which relates to many experimentally measurements, it is highly desirable to compute the free energy of the system on hand. The free energy is directly related to the probability of the system to be in a specific state. Absolute free energies are

almost impossible to calculate as the partition function is unknown for all but the simplest of cases.

$$F = -\frac{1}{\beta} \ln Q \,.$$

However the relative free energies of two states A and B can be calculated as

$$\Delta F_{AB} = F_B - F_A = -\beta^{-1} \ln \frac{Q_B}{Q_A} \,,$$

with either different Hamiltonians H_A and H_B or simply two different conformations of the same system [30]. The fraction of these partition functions can actually be computed by sampling.

It is often impossible to sample all states which contribute to a specific reaction by just sampling their densities according to their Boltzmann weights. The exponential nature of the Boltzmann factor leads to increasingly high computing times necessary to sample transition barriers. The problem can be mitigated by two types of approaches. Regions of the phase space which have not been sampled well enough are either known or unknown. If these regions are known, the Hamiltonian can be changed in a way to sample those regions better. However, if these regions are not known, more general method have to be used.

In this case, the dynamics of the system can be enhanced without changing the Hamiltonian. Another approach is to deform the energy surface and flatten it so barriers can more easily be overcome. Additionally, forces can be changed, the number of degrees of freedom be reduced or multi-copy approaches can be taken [30].

Here, some methods where the transition states and the problematic under-sampled regions will be explained. Replica exchange MD will be explained as one of the multi-copy approaches.

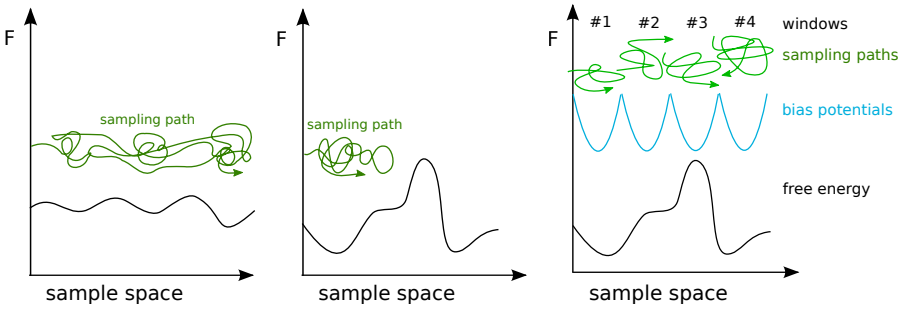
2.12.1 *Thermodynamic Integration - TI*

To compare the free energy of two different states A and B, the method of thermodynamic integration (TI) can be used. These two states can be two systems with different Hamiltonian functions altogether. Specifically, the Hamiltonian energies H_A and H_B have different dependencies on the spatial coordinates of the system. To calculate the free energy difference of these two states is not trivial as calculating differences of potential energies. To calculate the free energy of a given state, all contributing terms have to be summed over the phase space according to their Boltzmann weights. Thereby, entropic terms are considered in the partition function states A and B.

To calculate the difference in free energy of such two states, the parameter λ can be defined to smoothly transition from Hamiltonian H_A to H_B. The difference is then calculated by integrating along λ over the derivatives of H in respect to λ.

The simplicity of the method is one of its strengths. For many problems, thermodynamic cycles can be used in order to formulate the problem in a way that it can be solved with Thermodynamic Integration (TI). Hereby, computationally expensive transitions are replaced by multiple simpler ones[5].

2.12.2 *Umbrella Sampling - US*



(a) MD simulations can easily sample phase space with low free energy barriers.

(b) If the energy landscape contains high barriers, the MD simulation does not sufficiently sample the configuration space.

(c) The same free energy barrier as in Figure 2.3b is overcome by using multiple bias potentials. These potentials resemble horizontally flipped umbrellas covering a crowd of people, thus giving the method its name. The sampling paths for the different windows are mostly bound to their windows.

Figure 2.3: Sampling and barriers. Overcoming these barriers by Umbrella Sampling.

In many systems, two important conformational states can be separated by high energy barriers. These barriers prevent the system to cross from one state to the other in the proposed simulation time (shown in Figure 2.3b). As the probability of crossing such barriers scales exponentially with the negative of the potential as in the Boltzmann distribution, the time to cross such a barrier might be substantially higher than the simulation time. Examples include protein folding with transition times in the order of micro-seconds to seconds.

5 More details can be found in the review papers of Pohorille, Jarzynski, and Chipot [137] and Christ [30].

Enhanced sampling method such as Umbrella Sampling can mitigate this problem. Torrie and Valleau suggested in 1977 that by using arbitrary sampling distributions specifically to enhance estimates of free energy differences. They successfully tested the method on a simple Lennard-Jones fluid and measured the free energy of the gas-liquid phase transition [165].

Estimates of free energies between two states of a system are biased with additional potentials to increase the sampling rate of such transition states. After the biased simulation has been performed, the effect of the bias has to be reverted in order to calculate the wanted properties of the unbiased system [13]. In some cases, it might be possible to use a single biased simulation, and re-weighting the distributions with the ratio of the unbiased distribution function and the biased distribution function [13]. In many cases, the appropriate ratio is not known and, therefore, multiple simulations with different bias functions may be used. A particular useful scheme is the Umbrella Sampling scheme. It also has the advantage of being easily parallelizable.

Torrie and Valleau noted, that instead of using a single biasing function, it is often more convenient to use overlapping distributions, depicted in Figure 2.3, thus coining the name Umbrella Sampling. Here the same free energy barrier as in Figure 2.3b is overcome by using multiple bias potentials. Then, the sampling paths for the different windows are mostly bound to their windows with some overlap. It is very important that this overlap is not minuscule as is explained later in Section 2.12.4.

2.12.3 *Calculating the Potential of Mean Force*

The Potential of mean force (PMF) is the potential of the force of a given system keeping some coordinates fixed and averaging over all remaining configurations. This is a generalization of the terminology first introduced in 1935 by Kirkwood [88]. The PMF can be understood as the potential describing the dynamics as if the motion is on a free energy surface [29].

As stated by Roux, the PMF $W(\xi)$ along a reaction coordinate ξ is defined from the average distribution function $\rho(\xi)$ as

$$W(\xi) = W(\xi^*) - k_B T \ln \left[\frac{\langle \rho(\xi) \rangle}{\langle \rho(\xi^*) \rangle} \right] , \tag{2.2}$$

with ξ^* and $W(\xi^*)$ being arbitrary constants [142]. The average distribution function is obtained as

$$\langle \rho(\xi) \rangle = \frac{\int d\mathbf{R} \delta \left(\xi(\mathbf{R}) - \xi \right) e^{U(\mathbf{R})/k_B T}}{\int d\mathbf{R} e^{U(\mathbf{R})/k_B T}} . \tag{2.3}$$

$U(\mathbf{R})$ is the total internal energy of the system as a function of the coordinates $\mathbf{R}$. $\xi(\mathbf{R})$ expresses the functional dependency of the previously introduced reaction coordinate ξ in terms of all Cartesian coordinates $\mathbf{R}$ of the system. ξ can dependent in a simple manner on few coordinates of the system if it only depends on the coordinates of few atoms, such as in an angle between two bonds, or in a more complicated manner for a coordinate such as the Root-mean-square deviation (RMSD) Section 2.13.2.

As stated before, it is difficult to calculate the free energy and the PMF from standard MD simulations directly. Umbrella Sampling simulations present one of the most common ways of calculating the PMF as presented in the following.

Umbrella Sampling potentials $w_i(\xi)$ for the different windows i are introduced. For these, a common approach is to use harmonic potentials such as

$$w_i(\xi) = \frac{1}{2} K (\xi - \xi_i)^2 ,$$

with K being the force constant and ξ_i the central point of the potential as presented in Figure 2.3c. The un-biasing and combination of these multiple Umbrella Sampling windows is critical. With the average as defined in Equation 2.3, the biased distribution function for the i-th window is given as

$$\langle \rho(\xi) \rangle_{(i)} = e^{-w_i(\xi)/k_B T} \langle \rho(\xi) \rangle \langle e^{-w_i(\xi)/k_B T} \rangle^{-1} . \tag{2.4}$$

To un-bias the PMF as defined in Equation 2.2, the biasing potential has to be subtracted and the free energy F_i, which is associated with the potential, added [142].

$$W_i(\xi) = W(\xi^*) - k_B T \ln \left[\frac{\langle \rho(\xi) \rangle_{(i)}}{\langle \rho(\xi^*) \rangle} \right] - w_i(\xi) + F_i .$$

The free energy F_i is related to the potential as the simple average of the effect of the introduced biasing potential as

$$e^{-F_i/k_B T} = \langle e^{-w_i(\xi)/k_B T} \rangle . \tag{2.5}$$

The simplest approach to calculate F_i is to adjust the PMF functions until the associated F_i and F_j match for the adjacent windows i and j. The Weighted Histogram Analysis Method (WHAM) has proven to be more powerful as it uses all of the simulated data.

2.12.4 *Weighted Histogram Analysis Method*

The currently used WHAM equations are extensions by Kumar et al. [92] of the Multiple Histogram equations developed by Ferrenberg

and Swendsen [44, 45]. The following explanation of the WHAM equations is done accordingly to the explanation of Roux as it is most easily understood [142].

For N_W biased Umbrella Sampling simulations, the unbiased distribution function is the summation over all individual unbiased distribution functions multiplied by

$$\langle \rho(\xi) \rangle = \sum_{i=1}^{N_W} [\langle \rho(\xi) \rangle]_{(i)}^{\text{unbiased}} \left[\frac{n_i e^{-[w_i(\xi)-F_i]/k_B T}}{\sum_{j=1}^{N_W} n_j e^{-[w_j(\xi)-F_j]/k_B T}} \right], \tag{2.6}$$

with n_i being the number of independent snapshots taken for the i-th simulation [44, 92, 142].

According to the equations Equation 2.4 and Equation 2.5, the distribution function for each umbrella window can be unbiased as

$$[\langle \rho(\xi) \rangle]_{(i)}^{\text{unbiased}} = e^{w_i(\xi)/k_B T} \langle \rho(\xi) \rangle_{(i)} e^{-F_i(\xi)/k_B T}, \tag{2.7}$$

with

$$\langle \rho(\xi) \rangle = \sum_{i=1}^{N_W} n_i \langle \rho(\xi) \rangle_{(i)} \left[\sum_{j=1}^{N_W} n_j e^{-[w_j(\xi)-F_j]/k_B T} \right]^{-1}, \tag{2.8}$$

and

$$e^{-F_i/k_B T} = \int d\xi e^{-w_i/k_B T} \langle \rho(\xi) \rangle . \tag{2.9}$$

The equations Equation 2.8 and Equation 2.9 depend on one other and must be solved consistently. The simplest approach is to iteratively solve the equations in alternating succession. The first guess for the F_i can be chosen arbitrarily but better guesses result in faster convergence of the iteration procedure.

2.12.5 *Replica Exchange Molecular Dynamics*

The method of improving Monte Carlo simulations by having multiple copies of the system was independently developed by Berg, Billoire, and Foerster and Swendsen and Wang [15, 161]. Zhou extended the formalism to the method of Molecular Mechanics simulations [187]. In particular, multiple copies of the system, so called replicas, are run at different temperatures. These copies of the system do not interact continuously but sporadically exchange their temperatures under specific conditions. For a canonical ensemble, the Hamiltonian is weighted by the Boltzmann factor.

In replica exchange MD, a set of M temperatures and the same number of replicas is set. As the replicas can switch their temperatures, the labeling of the replicas is a permutation of the

temperatures. Each temperature only appears once. In the process of exchanging the replicas, the exchange of the coordinates of the replicas is equivalent to the exchange of their temperatures. While exchanging velocities, they are rescaled according to their new temperature to satisfy the condition of average kinetic energy. A rescaling by the square-root of the temperature quotient is therefore needed.

The detailed balanced condition is the principle of Markov chains that it cannot be distinguished whether a process is forward or backwards going at equilibrium. This principle was introduced by Ludwig Boltzmann in 1872 for collisions. In such processes the forward and backward rates are equal. From the detailed balanced condition it can be shown that the quotient of exchange rate of two replicas forwards and backwards is

$$\frac{w(X \to X')}{w(X' \to X)} = e^{-\Delta},$$

with

$$\Delta = (-\beta_n - \beta_m)\left(E(\mathbf{q}^{[i]}) - E(\mathbf{q}^{[j]})\right),$$

for replicas i and j with and their coordinates $\mathbf{q}^{[i]}$ and $\mathbf{q}^{[j]}$ at temperature T_n and T_m. One needs to distinguish between the indices of the replicas and temperatures as they are permuted from previous exchanges. The standard Metropolis criterion satisfies the condition mentioned above:

$$w(X \to X') = w(x_m^{[i]}|x_n^{[j]}) = \begin{cases} 1, & \text{if } \Delta \leqslant 0 \\ e^{-\Delta}, & \text{if } \Delta > 0\,. \end{cases} \tag{2.10}$$

The steps in Replica Exchange Molecular Dynamics (REMD) are then the following tow: First, the simulation of the replicas simultaneously and independently for a specific number of MD steps. Second, the exchange according to the Metropolis criterion (Equation 2.10) of adjacent replicas. REMD has proven to be one of the most successful and versatile techniques in the field of MD.

Hamiltonian Replica Exchange Molecular Dynamics

Hamiltonian Replica Exchange Molecular Dynamics (HREMD) is a generalization of the classic temperature replica exchange MD technique[6]. Instead of having M different temperatures T_m, a set of different m different Hamiltonian functions is defined for the system.

6 For a comprehensive review of Generalized-ensemble algorithms such as HREMD refer to [123].

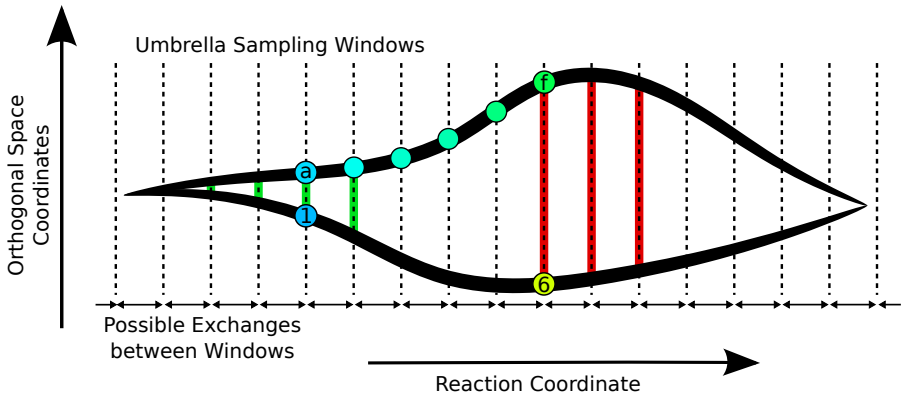


Figure 2.4: Advantages of REMD: The horizontal dimension represents the reaction coordinate whereas the vertical dimension represents a one-dimensional picture of the space orthogonal to the reaction coordinate. Assuming that the initialization procedure produced starting structures along the bottom pathway, conformations like (f) could not be sampled as the barrier (shown in red) is to high in this example. As the barrier is smaller on the left (depicted in green) conformations like (a) could be sampled. Using REMD these conformations can exchange all the way to windows further on the right. This way also the upper pathway can be sampled.

These Hamiltonian usually differ through added potentials which improve the sampling.

One approach is similar to temperature replica exchange in that the unperturbed system can be described as the lowest replica in direct comparison to the replica with standard temperature. Higher replicas are perturbed with increasingly higher potentials to improve the sampling speed. Another approach is to combine Umbrella Sampling (US) and REMD. Here, each Umbrella window represent one replica with a specific Hamiltonian defined by the standard Hamiltonian plus the Umbrella potential. The exchange of the systems of these replicas can increase sampling and mitigate specific problem with orthogonal sampling as describes later in Chapter .

Using HREMD, the sampling of the space orthogonal to the reaction coordinate could be improved as pathways along the reaction coordinate with different orthogonal components can also be sampled (see also Figure 2.4 for further explanation). The problem of insufficient sampling of orthogonal coordinates has been thoroughly investigated by [42] [42, 188].

2.12.6 *Theory of Reaction Rates*

For simple unrestrained MD simulations, reaction rates from one molecular conformation to another can be calculated directly. However, for many enhanced types of MD simulation reaction rates cannot be analyzed and calculated. The focus of most enhanced

simulations is the calculation of the free energy differences of the before-mentioned conformations. From the free energy alone only the distribution or the probabilities of the states can be calculated. The kinetics of the system are still hidden. As shown by Hummer, estimates of the local diffusion can be calculated from US simulations. Combining the local diffusion with the free energy can yield estimates of the reaction rates from different states of the system to one other [73].

Reactions along a one-dimensional reaction coordinate can be described by the one-dimensional Smoluchowski equation as explained by Hänggi and Borkovec [65]

$$\frac{\partial}{\partial t}\Psi(\xi, t) = \frac{\partial}{\partial \xi} D(\xi) e^{-\beta F(\xi)} \frac{\partial}{\partial \xi} \Psi(\xi, t) e^{\beta F(\xi)} .$$

Here the function $\Psi(\xi, t)$ describes the probability density along the one-dimensional coordinate ξ at time t. The free energy profile $F(\xi)$ is the one calculated by US simulations before. The following equation relates the local diffusion $D(\xi)$ along the one-dimensional reaction coordinate ξ to the time evolution of a specific Umbrella window and is accurate as long as the Umbrella windows are small enough [73, 189]

$$D(\xi) = \frac{\sigma^2(\xi)}{\tau} .$$

The auto-correlation time can be calculated using the method of Hess [68] which uses block averaging and a double exponential decay fit of the correlation function. The diffusion perpendicular to the reaction coordinate ξ influences the local kinetics and thus the local diffusion coefficient

The mean first passage time is the inverse of the specific reaction rate. It can be calculated as a double integral along the one-dimensional reaction coordinate. For see derivation refer please refer to [9, 162, 177, 190]. The first boundary was set to be reflecting and the second boundary to be absorbing. This assumption will be satisfied for flipping processes from the intra-helical to extra-helical state in a flipping of a base around the helical axis of the DNA. States in the intra-helical state are quite stable and states in the extra-helical state undergo the subsequent direct repair process. They are therefore taken out of the system as such.

The mean first passage time acMFPT from the first state A the second state B is thus given as

$$\tau_+ = \tau_{A\to B} = \int_{\xi_A}^{\xi_B} d\xi' \frac{e^{\beta F(\xi')}}{D(\xi')} \int_{\xi_A}^{\xi'} d\xi'' e^{-\beta F(\xi'')} .$$

Similarly, the mean first passage time of the inverse process can be calculated as:

$$\tau_- = \tau_{B\to A} = \int_{\xi_A}^{\xi_B} d\xi' \frac{e^{\beta F(\xi')}}{D(\xi')} \int_{\xi'}^{\xi_B} d\xi'' e^{-\beta F(\xi'')} ,$$

where the probability is integrated along the backwards direction. The rates are related as

$$k_+ = \frac{1}{\tau_+} ; k_- = \frac{1}{\tau_-} ; k_{EQ} = \frac{k_+}{k_-} .$$

The equilibrium rate k_{EQ} depends purely on the free energy difference of the two states

$$\Delta F = \beta^{-1} \ln (k_{EQ}) ,$$

and can be double checked with the values calculated from the reaction rates.

2.13 ANALYSIS

2.13.1 *Measuring Observables*

Evaluating macroscopic observables from a single or just a few simulations is possible due to the ergodic theorem Table 2.10. As such, time averages are equivalent to ensemble averages for sufficiently long simulation times. This is allows for the measurement of macroscopic observables as temperature which can also be measured in experiments.

2.13.2 *Root Mean Square Deviation - RMSD*

One of the most useful and therefore most commonly used measurements in the analysis of MD-simulations is the measure of RMSD. It is the measure of the average distance between the atoms of superimposed structures of different molecules or the same molecules in different states. Therefore the similarity of different conformations can be measured.

Typically, RMSDs are calculated in respect to the backbone atoms of proteins. Sometimes, an all-heavy atom RMSD is calculated. For DNA, the RMSD of the nucleic backbone gives a good impression of the deviation of the backbone and overall structure.

The RSMD for two sets $\overline{\mathbf{v}} = \{\mathbf{v_i}, .., \mathbf{v_n}\}$ and $\overline{\mathbf{w}} = \{\mathbf{w_i}, .., \mathbf{w_n}\}$ with n atoms each, is defined[7] as

$$\mathrm{RMSD}(\overline{\mathbf{v}}, \overline{\mathbf{w}}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{v_i} - \mathbf{w_i}\|^2} \tag{2.11}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n} ((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2)}\,. \tag{2.12}$$

The best-fit RMSD is then calculated by optimizing the translation and rotation such that the RMSD is minimal. The needed translation and rotation can be calculated by using quaternions proposed by Coutsias, Seok, and Dill [33] or similarly by the Kabsch-algorithm by Kabsch [82].

Typical RMSD values depend on the system and the given problem. As X-ray resolution is in the order of 2 Å–3.5 Å, fitting a model below this value gives already a good solution to the fitting problem of X-ray scattering. For docking two ligands, an heavy-atom RMSD of 2 Å refers to a good docked solution. Solutions below 1 Å are perfect to the point of atomic fluctuations. In this manner, conformations in the range of 1 Å are considered to belong to the same conformation.

2.13.3 *DNA Parameters Analysis*

The scientific community decided to define parameters to describe the geometry of nucleic acid structures in 1988 at the EMBO Workshop on DNA Curvature and Bending [39]. The parameters depicted in Figure 2.5 have been used to describe the rigid-body parameters of base pairs and base pair steps ever since. Some of these parameters will be used to describe specific local and global parameters of the studied DNA structures. Global properties such as bending can often be described by a set of local properties such as the roll angle.

2.14 METHODS

Some supplemental methods used over the course of this thesis are described in the following sections.

2.14.1 *Restraints*

Restraints are one of the most important tools in MD simulations. They can be used as one of the tools in the setup chain or as a tool in enhanced sampling methods. Different restraints on the

7 Refer to [36] for details.

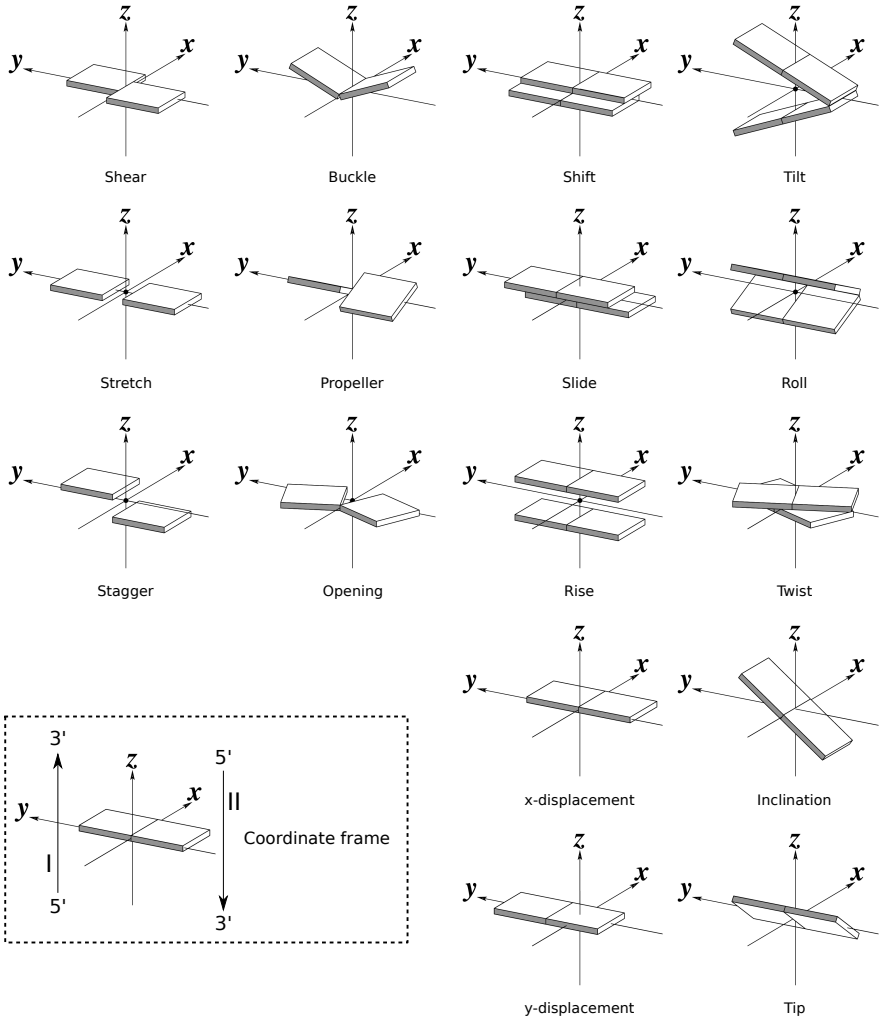


Figure 2.5: Visual representation of rigid-body nucleic parameters which describe the geometry of base pairs and base pair steps. The reference frame is depicted at the lower left. Figure analogously to [107, 108].

coordinates of the system can be used. These restraints act in some form or another on the atomic positions in the system. Other types of restraints are possible but far more difficult to use and implement.

Restraints generally induce some kind of energy penalty $U_{\text{restraint}}$. By calculating the gradient along all coordinates x, y, and z, the forces

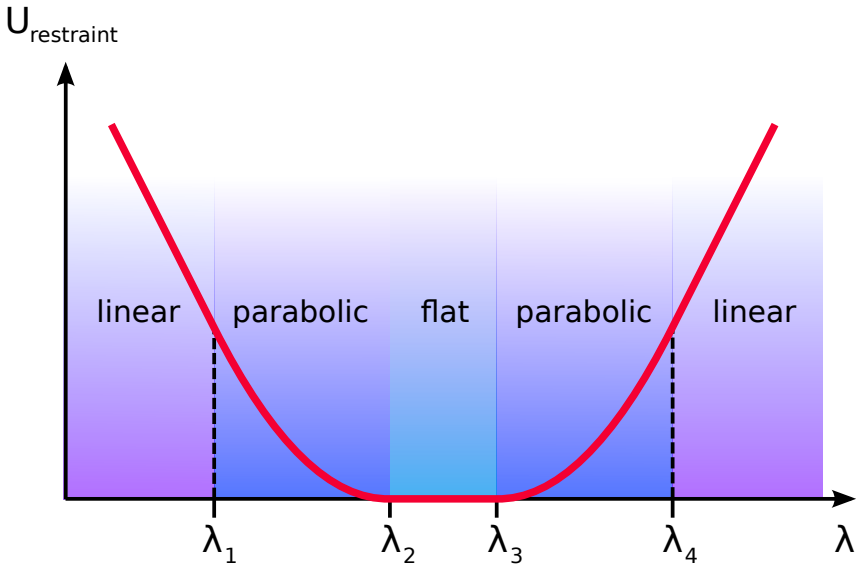


Figure 2.6: Standard restraint potential.

are calculated as $\mathbf{F}_{\text{restraint}} = \nabla U_{\text{restraint}}$. The most commonly used form of penalty functions is

$$U_{\text{restraint}}(\lambda) = \begin{cases} 2k_2(\lambda_1 - \lambda_2)\lambda - k_2\left(\lambda_1^2 - \lambda_2^2\right), & \text{if } \lambda < \lambda_1 \\ k_2(\lambda - \lambda_2)^2, & \text{if } \lambda_1 \leqslant \lambda < \lambda_2 \\ 0, & \text{if } \lambda_2 \leqslant \lambda < \lambda_3 \\ k_3(\lambda - \lambda_3)^2, & \text{if } \lambda_3 \leqslant \lambda < \lambda_4 \\ 2k_3(\lambda_4 - \lambda_3)\lambda - k_3\left(\lambda_4^2 - \lambda_3^2\right), & \text{if } \lambda_4 \leqslant \lambda \end{cases}$$

A sketch of such a function with $k_2 = k_3$ is shown in Figure 2.6. For the use of Umbrella Sampling, a standard parabola is required. Thus, by setting $k_2 = k_3$, $\lambda_2 = \lambda_3$, and λ_1 and λ_4 so far away from λ_2, these values cannot be reached by λ due to very high restraint potentials.

The commonly used restraints are restraints on positions, angles, dihedral angles, distances, RMSD and Distance-RMSD (D-RMSD). Positional restraints are used in the equilibration process to heat up the solvent without changing the conformation of the solute too drastically. Additionally, for the creation of specific structures, restraints on the positions of all or a subset of atoms can be used. Thereby, structures which are similar in some aspects to one structures but other aspects to another structure, can be created.

Further, instead of using single atoms, the optionally mass-weighted center of masses of groups of atoms can be used. This was not possible in the Particle Mesh Ewald Molecular Dynamics (PMEMD) implementation in AMBER for dihedral restraints and was implemented for my purposes inside the AMBER12 code specifically.

The details of the dihedral and RMSD restraints used later will be explained in the related chapters.

3

MOTIVATION AND MODEL

3.1 QUALITATIVE MODEL

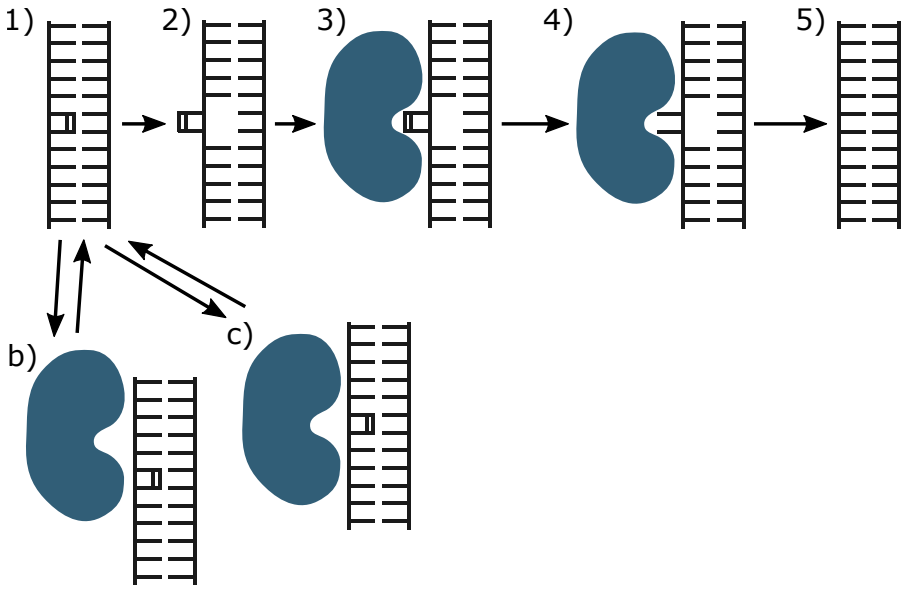


Figure 3.1: Model 1: Passive Recognition In the passive recognition process, the repair protein completely relies on the spontaneous flipping transition of the damaged base dimer. 1) The DNA is in the damaged state and the base is intra-helical. b) If the protein attaches in the vicinity of only undamaged sites, nothing happens. c) If the protein attaches to the damaged site in the intra-helical state, no repair will be done in this mechanism. If however, the damaged base spontaneously flips into the extra-helical state (2)) and the protein attaches then to the damaged extra-helical site (3)), the splitting of the dimer can be performed (4)). The protein detaches afterwards (5)).

In the following, some of the most likely damage recognition and repair mechanisms will be presented and discussed. Specifically, a passive and an active damage recognition during the repair process have been proposed [134]. The terms of conformational selection and induced fit do not describe the mechanism of DNA repair sufficiently as they are defined for the process of protein ligand binding. However, in a general sense, they can help explain and differentiate the hypothetical DNA repair mechanisms. Generally, the change of the protein structure in binding with a ligand can occur before (conformational selection) or after (induced fit) the association with the ligand [57]. Here, this does not refer to the conformational

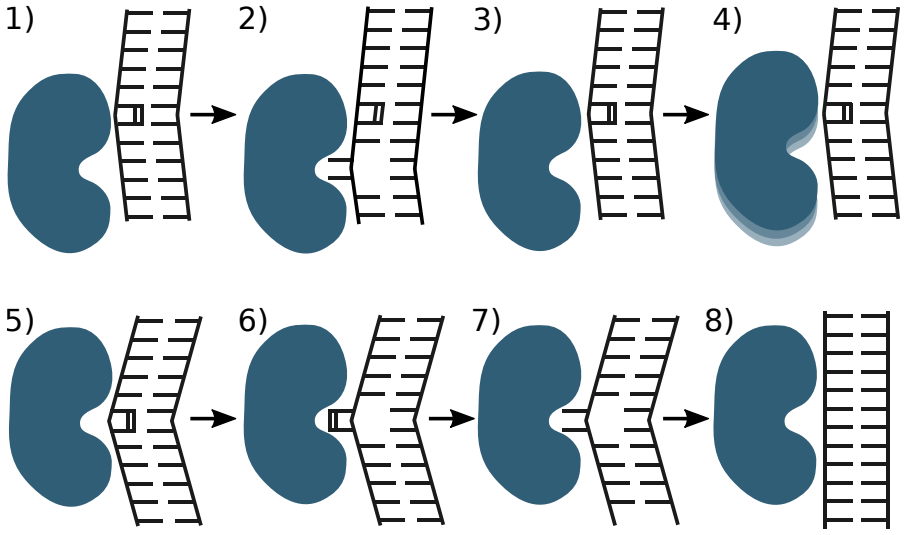


Figure 3.2: Model 2: Recognition by flipping of every base: 1) The protein attaches to any site and checks the bases whether they need repair or not by flipping the bases (2). After the bases are in the intra-helical configuration (3), the protein slides in uncontrolled direction (one-dimensional random walk) to the next base (4). When the protein reaches the damaged bases (5), it again flips them into the extra-helical conformation (6) and repairs them (7). Afterwards, the repaired bases flip back into the intra-helical position (8). Here, all bases are flipped out and checked. For the sake of simplicity only the sliding (and no hopping) along the DNA is depicted. In the native state the DNA is bend at the site of the lesion. During binding the DNA changes its local conformation at the checked site.

change of the repair protein but to the conformational change of the DNA in the binding process with the repair protein.

Therefore, a first model is therefore that the damaged DNA flips into the extra-helical state before binding with the repair protein. This could be described as the passive process in which the repair enzyme relies on the random base flipping motion to bind to the flipped out damaged DNA base. The detailed steps involved are explained in Figure 3.1.

The second model is that the damaged DNA does not change its structure before binding to the repair protein. As the protein cannot detect changes of the internal, i.e. visible from the outside, structure, every base would be completely or partially flipped out in order for the protein to repair the damaged bases (see Figure 3.2).

As a third model, it is also possible that only damaged bases undergo the flipping transition. In this case a mechanism for the differentiation before the flipping process is required. Likely, the protein increases the binding affinity of the DNA at the damaged site recognizing the already present changes in the structure. By relying on the fact that the transition from the conformation in bulk into the proper bound form happens more readily, mostly only damaged

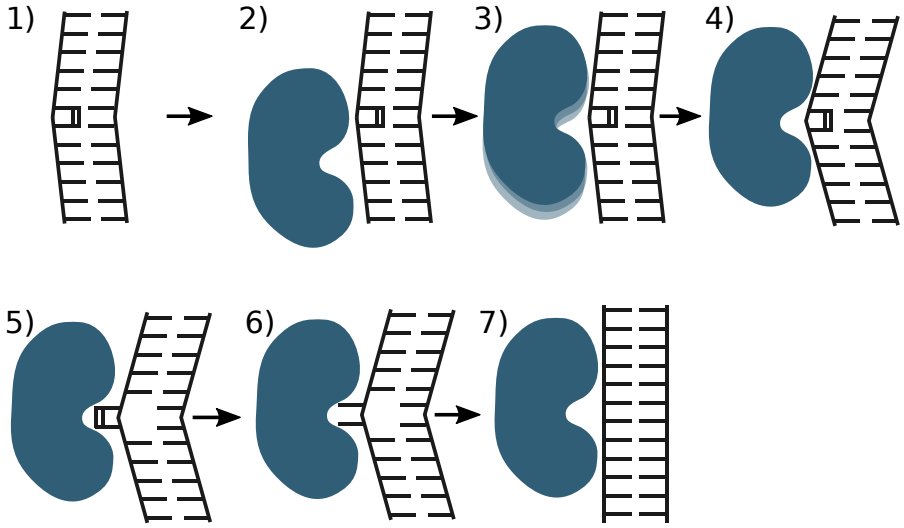


Figure 3.3: Model 3: Recognition by attaching closely to damaged bases: The DNA is, if damaged, in a slightly bend configuration (1). If the protein reaches an undamaged site (2), it continues to slide as it does not bind properly to the DNA. When the protein reaches the damaged site by sliding (one-dimensional random walk) (3), it attaches more closely to the DNA (4) by changing into an even stronger bend conformation. Because the protein and DNA are in close contact, the damaged base is flipped into the extra-helical configuration (5). After repair (6), the thymine bases flip again into the intra-helical configuration (7) and the DNA takes on the native B-DNA configuration.

bases would flip into the extra-helical repair site of the protein. For this process see Figure 3.3.

It is possible to simulate some sub-processes of these repair mechanisms using current MD techniques. Yet, the complete process cannot be currently simulated as a whole.

3.2 QUANTITATIVE MODEL

From heron, a quantitative model of the sub-processes involved in the hypotheses will be given. The three hypotheses will now be looked at in further detail quantitatively. It is of interest to calculate the maximum free energy differences required for these involved sub-processes to work. The following assumption will be taken into account. The human genome and DNA damage repair by glycosylases are used as an analogue system. For the studied system of photolyases in E. coli. not all the required data is available. The human genome has a size of 7×10^9 bases. With 10^5 DNA repair glycosylases present in one nucleus, one repair enzyme effectively surveys 7×10^4 base pairs. Here, it was assumed that all base-pairs are equally accessible which might not be the case in highly packed nucleosome structures. Nonetheless, this calculation will present a

lower bound for the rate of damage repair necessary. 10^4 damages accumulate every day per cell [53]. For fast replicating cells in the human body, the damage needs to be repaired in the time-span of a single day as this is the duration of a whole cell cycle.

3.2.1 *Three-dimensional search*

The first approximations for the search time of a damaged base can be made using the Smoluchowski diffusion equation [151]. Here, a standard three dimensional search is assumed. To locate a site of radius r in a volume V with the diffusion coefficient D_3, the enzyme searches on average for the time of

$$t_{\text{search,3D}} = {}^{V}/_{D_3 r}\,.$$

The time is directly proportional to the volume which each enzyme needs to sample. For the sake of this calculation it does not matter if one considers N enzymes in the total volume V_{nucleus} or one enzyme in its responsible volume $V_{\text{enzyme-res}} = {}^{V_{\text{nucleus}}}/_{N}$. The time to search the volume decreases with increasing rates of three-dimensional diffusion D_3 and increasing target radius r. That is, a larger target site can be found more easily. Experimental measurements give value of $1 \times 10^8\,\text{nm}^2/s$ for typical enzyme of a diameter of 50 Å or 5 nm [53]. The nuclear volume is around $1 \times 10^{11}\,\text{nm}^3$ and the target radius of a single base can be approximated as $0.17\,\text{nm}$ as the bases are $0.34\,\text{nm}$ apart along the DNA strands. For just one repair enzyme its responsible volume is $1 \times 10^6\,\text{nm}^3$ and the time for searching a damaged base pair subsequently only 2.9×10^{-2} s or $29\,\text{ms}$. This is obviously much shorter than the duration of DNA replication. Nonetheless, it was completely neglected that the enzyme might stay at any sites and interrogate them. The model completely breaks down if the enzyme stops its three-dimensional free diffusion at any point. Firstly, it will be blocked by many obstacles such as the DNA itself. Secondly, by transitioning into an interrogation complex with the DNA, it will stay at such sites according to the binding coefficient and the interrogation life-time $k_{\textbf{off}}$. Yet, for the first model the current approach can be seen as a valid assumption. Thus, for the first hypothesis this model is investigated further in the following Section 3.2.2.

3.2.2 *Three dimensional diffusion and Passive Recognition*

For three-dimensional, free diffusion it must be assumed that the enzyme does not stop at non-damaged base-pair sites. As the structure of damaged and undamaged base pairs is not very different

as seen from the outside of the DNA, the difference is negligible if the damaged base (or dimer) is in the intra-helical position.

Therefore, a damaged base must be recognized in its extra-helical conformation. As there are on average only less than one damaged sites present per volume which one enzyme covers on average, the search time is not changed by the damaged bases. Until the enzyme finds the damaged base, it completely diffuses freely. If the damaged base would always be in the extra-helical position, it could be found in the search time of $t_{search,3D} = 29\,ms$. This is however not the case. The total search time is thus

$$t_{search,total,3d} = t_{search,3D} \cdot n_{visits}\,, s$$

where n_{visits} describes the number of visits needed to find the base in the extra-helical configuration. To repair the damage in one day, it therefore needs to find the DNA in at least one of 2.98×10^6 visits. This probability is related to the free-energy difference $\Delta F = F_{out} - F_{in}$ as

$$\frac{P(out)}{P(in)} = e^{-\beta \Delta F}\,.$$

$P(in)$ is approximately 1 in this calculation. The free energy difference between the extra-helical and the intra-helical state must therefore be smaller than

$$\Delta F = -\beta^{-1} \ln P(out)\,.$$

For the values stated above the difference of the free energy of the extra-helical state and the intra-helical state cannot be larger than $8.88\,kcal/mol$. As this calculation relies on many simplistic assumptions, a more rigorous calculation including one-dimensional search will be introduced.

To further elucidate this model, one can now calculate how an error in the assumptions effects the predicted free energy. For the current model, the resulting free energy changes by $-\beta^{-1} \ln(10) \approx 1.37$ if the probability increases by a factor of 10. Vice versa, the free energy decreases by only 1.37 if the probability decreases 10-fold.

3.2.3 *Sliding and Hopping*

As explained above, three-dimensional diffusion alone is sufficient to lead to rapid encounters of the enzyme with damaged sites. However, in real systems the search is limited by the fact that the repair protein must attach and test whether base pairs are damaged or not. The attachment and dissociation into the bulk slows down the search process dramatically. To increase the speed of search it is highly advantageous if the repair enzyme does not detach too

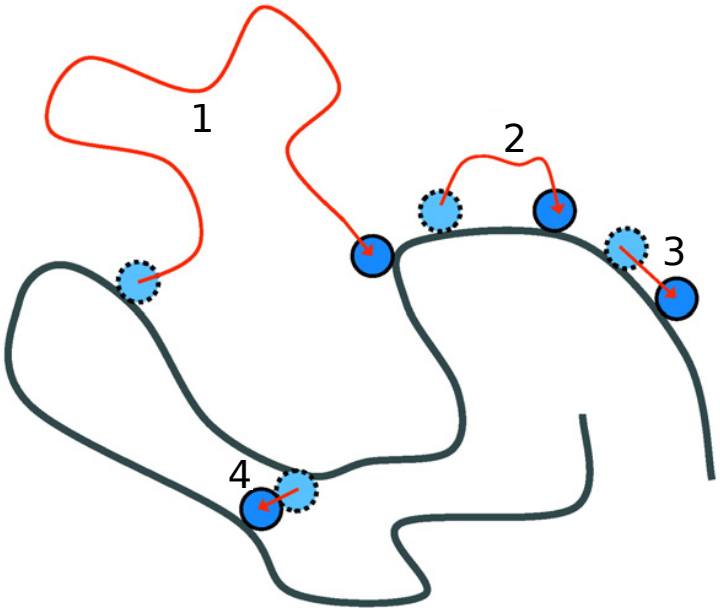


Figure 3.4: Illustration of the four different mechanisms by which a protein searches DNA: three-dimensional diffusion (1); hopping or correlated transfer (2); one-dimensional sliding (3); inter-segmental transfer (4) [60].

quickly but slides along the DNA for some time before dissociating into the bulk. The sliding along the DNA can be described by a one-dimensional search to the target site. Such a search process involves multiple mechanisms. As shown in Figure 3.4 multiple mechanisms play a part: Three-dimensional diffusion (1), hopping for only short distances (2), sliding along the DNA (3) and hopping to other parts of the DNA, i.e. inter-strand hopping (4). For simplicity, the system is explained in terms of the two mechanisms of three-dimensional and one-dimensional search as shown in Figure 3.5.

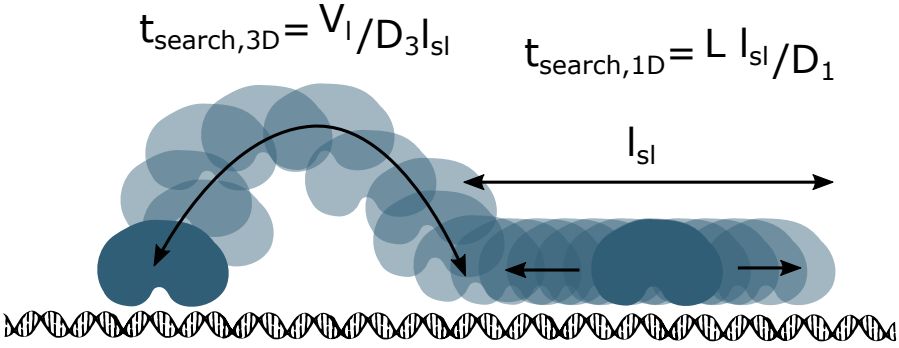


Figure 3.5: Here x denotes an arbitrary reaction coordinate.

As the motion along DNA is not directional, the enzyme moves according to the rules of one-dimensional random walk. Thus, the search time for sliding increases by the power of two with the number of steps, i.e. the number of base-pairs investigated. In most systems

the theoretical optimum sliding length l_{sl} of 10–100 base pair steps are in accordance with experimental values of the sliding length [53]. The protein has a limited duration for staying alongside the DNA. It is directly anti-proportional to the dissociation rate k_{off} as

$$t_{bound} = 1/k_{off}\,.$$

As the sliding along the DNA is described by a one-dimensional random walk, the sliding length can be expressed as $l_{sl} = \sqrt{D_1 t_{bound}}$. The total search time can then be calculated as

$$t_{search} = t_{search,3D} + t_{search,1D} = {}^{V}/_{D_3 r} + {}^{L l_{sl}}/_{D_1}\,.$$

From this equation, quantitative predictions can be made for the three previously explained models.

3.2.4 *Model for the flipping process*

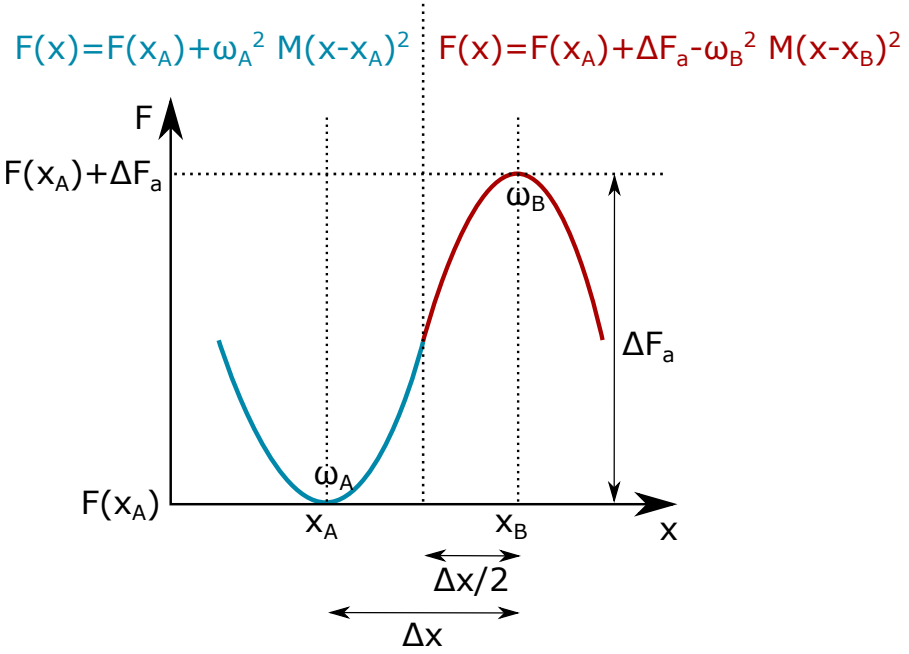


Figure 3.6

To relate the free energy barrier of the flipping process to the average time of visit of the protein at a specific DNA base, the following model will be used. It is calculated accordingly to the model of Hänggi and Borkovec [65]. Without any knowledge of the free energy potential, it is assumed that the potential is of harmonic form. Two regions are studied in detail.

The transition is explained in terms of a one-dimensional coordinate x. The start state A is at the coordinate $x = x_A$. The transition state B is located at $x = x_B$. For this calculation it is

sufficient to know that the final state C has a free energy which is lower than that of the transition state. The potential is equally split into two regions and given by

$$F(x) = \begin{cases} F(x_A) + \omega_A^2 M(x - x_A)^2, & \text{if } x \leqslant x + {}^{\Delta x}/_2 \\ F(x_A) + \Delta F_a - \omega_B^2 M(x - x_B)^2, & \text{if } x > x + {}^{\Delta x}/_2 \,. \end{cases} \tag{3.1}$$

If the region is split equally as shown in Figure 3.6, the strengths of the two harmonic potentials ω_A and ω_B are equal. It can then be calculated that

$$\omega^2 = \omega_A \cdot \omega_B = \omega_A^2 = \frac{2\Delta F}{\Delta x^2} \,.$$

In general, the rate between two states is calculated by the Arrhenius equation as

$$k = A \cdot exp(-\beta \Delta F_a) \,,$$

where ΔF_a describes the activation free energy as used above. The factor A can be temperature and free energy dependent [11]. For a simple transition state, however, it can be calculated using Kramers rate theory [65]. For the potential described above in Equation 3.1, the rate from state A to state C is then

$$k_{A \to C} = \frac{\omega_A \omega_B}{2\pi\gamma} exp(-\beta \Delta F_a) \,,$$

where γ denotes the damping relaxation rate. This can in turn be calculated from the drag coefficient for low Reynolds numbers which is the case here. The drag coefficient in this regime can be calculated as

$$\zeta = 6\pi\eta r \,.$$

In total, the rate is given by

$$k_{A \to C} = \frac{2\beta \Delta F_a D}{\pi \Delta x^2} exp(-\beta \Delta F_a) \,.$$

For the flipping process the length of the transition along the reaction coordinate x can be estimated to be in the magnitude of 2.5 nm or 25 Å for a rotational flip of 180° around an axis of radius 0.8 nm. The radius of the base around the flipping axis has been measured in the corresponding crystal structure.

The diffusion constant can be calculated for spherical particles as

$$D = \frac{k_B T}{6\pi\eta r} \,,$$

where η denotes the viscosity and r the radius of the spherical particle. The viscosity in water is approximately $0.798 \times 10^{-3}\,\mathrm{Ns/m^2}$. Now, the radius of the base has to be approximated. Here, the radius was given by the base at the widest position. A more complicated and accurate approximation could be used but the current approximation should suffice. The diffusion constant is then calculated to be roughly $1.38 \times 10^{-9}\,\mathrm{m^2/s}$.

The following equation has to be solved (numerically) for $\xi = \beta \Delta F_a$

$$\frac{k_{A\to C}}{C} = \xi \exp(-\xi)\,. \tag{3.2}$$

Here, the constant C is calculated as approximately $1.37 \times 10^9/\mathrm{s}$. The time of stay of the protein at a specific base site is related to the minimum complete repair rate at that site. As the chemical repair of splitting the dimer happens in the regime of fempto-seconds, the limiting step is the flipping rate. The flipping is in turn largely correlated to the free energy barrier of this process. Therefore, a concrete relation of the free energy barrier of the flipping and time of stay of the protein at one base is found.

3.2.5 *Model 2 Quantitatively*

For our analogue system, the repair enzyme stays at every base for $50\,\mu\mathrm{s}$ [53]. In this time, the whole repair process must be finished. Therefore, the rate must be larger than

$$k_{A\to C} = 2 \times 10^4/\mathrm{s}\,.$$

Equation 3.2 can now be solved for ΔF_a. Under the assumptions of Section 3.2.4 this gives a maximum free energy of $6.7\,\mathrm{kcal/mol}$ for the difference between the bulk intra-helical and the extra-helical bound state.

3.2.6 *Model 3 Quantitatively*

The third hypotheses builds upon the second model but the recognition mechanism is further enhanced for damaged bases. The upper boundary for the flipping of the undamaged bases remains the same. As the damaged DNA leads to an easier binding of the protein to the damaged site of the DNA, it can be assumed that the overall recognition and repair time is extended for the damaged site. Thus, if the protein can bind more easily to the damaged site, expressed by the binding difference ΔF_{bind} between undamaged and damaged sites, the time of visit of the protein at such sites is extended by the Boltzmann factor of such sites

$$t_{visit,damaged} = exp\left(-\beta F_{bind}\right) \cdot t_{visit,undamaged} \,. \tag{3.3}$$

Thus, the free energy for flipping is too high to allow the protein to stay long enough at every base that the undamaged base flips out in the presence of the repair enzyme. As the protein, however, stays longer at the damaged site according to Equation 3.3 this now increases the maximum free energy barrier for flipping by the F_{bind}.

3.3 TESTING MODELS USING MOLECULAR DYNAMICS

In addition to experiments, it is nowadays possible to study the structures, their conformations and their dynamics with MD simulations. In particular, the given models/hypotheses can be tested using multiple different MD approaches specific to the problem. Chapter 5 to Chapter 8 will give tangible answers to the given hypotheses.

The current approximations will be updated with measured local diffusion rates in later chapters of this thesis to get a better approximation for the possible free energy differences in the system.

4

EXPERIMENTAL AND COMPUTATIONAL LITERATURE ON DNA DAMAGE RECOGNITION AND REPAIR

4.1 EXPERIMENTAL LITERATURE

Experimental findings about properties of CPD DNA damage and repair are manifold. Studies on the structure are the most predominant. One of the first structural analysis was published in 1988 by Husain, Griffith, and Sancar . By using gel electrophoresis and quantitative electron microscopy it was shown that the existence of CPD lesions in DNA leads to a slower migration of the DNA fragments during electrophoresis in comparison to native DNA. This slow migration in the gel was explained by 30° bending of the DNA induced by the CPD damage [75]. The same method was used by Wang and Taylor to calculate a bending angle of approximately 7°. Due to these contradicting results, it could only be concluded that CPD causes some form of DNA bending.

This problem was later resolved by more accurate methods of structure analysis such as crystallography and NMR. Numerous experimental structure analysis have been performed in recent decades.

McAteer et al. resolved CPD containing duplex DNA and its native associated sequence by using NMR at 750 MHz [113]. They observed an intra-helical partially base-paired arrangement of the CPD bases within the DNA double helix. Additional to changes in local structure, helical structural changes were calculated in detail.

Lee, Choi, and Choi explained why CPD lesion only rarely (4%) lead to mutations during trans-lesion replication. The insertion of thymine residues opposite of the lesion is prohibited by structural distortions [97].

Lee et al. later found that the distortion in the DNA is increased if the CPD lesion leads to so called wobble pairs where WC partners of CPD thymines are replaced by one or two guanine bases. The change in overall structure is comparable to bending of DNA with 6-4PP lesions.

Significant damage-induced distortion of the helical DNA structure with respect to regular B-DNA in the absence of a bound repair enzyme (PDB:1N4E) [133] were measured by crystallography. Park et al. claim overall helical axis bending of 30° toward the major groove and unwinding of 9°.

Very importantly, first structures of extra-helical conformations in complex with photolyase were observed by Mees et al. The DNA structure contains a synthetic CPD analogue with the same cis-syn stereo-chemistry as natural CPD lesions. The intra-dimer phosphor di-ester was replaced by a methylene di-ether for sample preparation. A natural CPD would be repaired very quickly in photolyase containing solutions under small amounts of natural light which was prevented by the introduction of the CPD analogue [117]. This structure was used extensively as a starting structure during this thesis. For the simulations in the course of my thesis the CPD analogue was replaced by CPD or two thymines as appropriate.

Other recent crystal structures of photolyase complexes have found similar structural properties [87]. Kiontke et al. observed a specially large binding site for class II photolyases. In the external flipped configuration, water molecules play an important role in the recognition procedure.

At least as many observations have been made on 6-4PPs - the other main photo-induced type of damage. All NMR [98] and crystal structures [61, 62, 96, 148, 182] show that 6-4PP lesions have very strong global structural changes. Recognition likely happens by conformational selection of damaged DNA. As photolyase are very similar to their 6-4PP counterpart, a comparable recognition mechanism is likely applied for CPD containing DNA.

The absence of complexes of photolyase and DNA in the intra-helical configuration [179] does not clearly indicate a passive repair mechanism as stated by Wilson et al. However, the missing of extra-helical crystal structures without proteins indicates clearly that the extra-helical state is disfavored in comparison to the intra-helical state in bulk.

After binding of the damaged site to the repair protein, the subsequent looping out transition of the CPD into the enzyme active site can be considered as an induced fit step of the repair process. Indeed, for other repair processes, e.g. the recognition of oxidatively damaged guanine to 8-oxo-guanine, such encounter complexes with strongly bend DNA and importantly an intra-helical damaged 8-oxo-guanine have been structurally characterized [23].

A study using β-cyclodextrin to trap spontaneously flipped out bases shows that the flipping rate of $3.5 \times 10^{-3}/s$ is six to seven orders of magnitude slower than a comparative study in enzyme/DNA complexes [154]. This study suggests that a simple trapping mechanism of flipped out damaged bases is not a probable recognition mechanism for damaged DNA.

Typically, DNA repair enzymes recognize a damaged DNA-site with high affinity and specificity. For example, the E. coli DNA photolyase binds a thymine dimer containing DNA with a dissociation constant in the nano-molar regime ($\approx 30\,nM$) whereas

the binding to undamaged DNA of the same sequence is 7.5×10^3 times lower [76].

Due to the availability of crystal structures in the extra-helical configuration [87, 117], I chose the simple repair process of photolyases as our model system to understand the general repair mechanisms of CPD damage repair. The FAD cofactor plays an important role as the extra-helical state of the damaged site is required for the enzymatic repair by photolyases [146].

4.2 COMPUTATIONAL LITERATURE

The topic of recognition of DNA damages and DNA damage repair has been studied with the method of MD simulation for some time. The first extensive studies were done in the late 1990s. At that time it was possible to study the local conformations and structural effects but it was not possible to study most dynamical properties of DNA damage. Even now, it requires advanced molecular methods to study these properties of DNA repair.

General simulations started the study of the effect of CPD lesions on DNA. First short (500 ps) free molecular dynamics simulations on the structure of 1TTD - a cis,syn-CPD damaged DNA dodecamer - have shown slight changes in orientation around the glycosole bond for the 5′ thymine of CPD and a global curvature increase of 10° in comparison to native DNA. The results were in good agreement with NMR/Nuclear Overhauser Effect (NOE) measurements. For these short simulations, the hypothesis was made that many small distortions of the DNA by the lesions are responsible for the recognition mechanism rather than a global structural change [119]. Simulations by Spector, Cheatham, and Kollman in 1997 of the structural effects of cis-syn CPD dimer and 6-4 adducts as well as native DNA have been in good agreement with NMR analysis in terms of structural features such as bending angle. These simulations show that MD simulations are well suited for this type of analysis and that the model describes DNA and DNA lesions accurately enough to be useful. For a more detailed explanation of these studies refer to [153].

Apart from measuring the free energy profile of base flipping by US in the complex with cytosine-C5-methyiltransferase from HhaI DNA methyltransferase (HhaI), Huang and MacKerell measured the free energy of un-damaged cytosine base flipping. It is approximately 15 kcal/mol.

QM/MD approaches using Density Functional Theory (DFT) have been used to describe the region CPD bonds between the thymine bases. The bond between the C_5 atoms can be broken with little effort (barrier-less) whereas the free energy barrier for breaking the C6 bond has been calculated to have an upper limit of 2.5 kcal/mol

Type	Measurement of	Result	Reference
Free MD	Structure	Agreement w. experiments	[153]
Free MD	Structural changes	Global bending incrd. by CPD	[119]
US	C base flipping	$\Delta F = 15\,kcal/mol$	[71]
QM/MD	CPD bond break	$\Delta F = 2.5\,kcal/mol$	[112]
US	CPD base flipping	$\Delta F = 10\,kcal/mol$	[54]
US ncc.	CPD base flipping	ext. state $6.5\,kcal/mol >$ int. state	[128]
US ncc.	CPD base flipping	$\Delta F = 7.5\,kcal/mol$	[129]
US	Partner base flipping	$\Delta F = 10\,kcal/mol$	[186]

Table 4.1: Some of the referenced results obtained by MD simulations. Abbreviations used: not complete coverage (ncc.); extra-helical (ext.); intra-helical (int.); with (w.); increased (incrd.).

[112]. These types of simulation can give us an indication of the difference of CPD to two thymine bases chemically.

Further, impressive results have been obtained by the direct study of the base flipping reaction. O'Neil, Grossfield, and Wiest used a two-dimensional pseudo-dihedral coordinate to calculate the free energy barrier or PMF for the lesion-flipping process. Their results show that the external flipped state is only $6.5\,kcal/mol$ higher than the flipped-in state. This barrier height is not entirely conclusively in pointing to either a extra-helical (conformational selection) or intra-helical (induced fit) recognition mechanism [128]. By not doing a comparative study to undamaged bases, these results have to be compared to other studies using different methodology. Therefore, systematic error cannot be excluded.

A later study of O'Neil, Wiest, and O'Nei showed the dependence of the free energy on the adjacent base pairs. Adjacent G-C pairs (G adjacent to CPD) show a significantly higher base flipping free energy barrier height of $7.5\,kcal/mol$ [129]. Stacking interactions are quite likely the important factor for this difference.

Studies of base flipping Umbrella Sampling simulations for cis-syn CPD damaged DNA have shown a significant reduction in free energy barrier of $2.5\,kcal/mol$ in comparison to undamaged DNA. The final barrier height was measured as $10\,kcal/mol$ for CPD damaged DNA [54]. It may be noted that these results are interpolated from a limited region of the reaction coordinate in order to obtain the total free energy difference. Therefore, the resulting free energy barrier cannot be ultimately measured with as much precision in comparison to complete coverage of Umbrella Sampling simulations.

Other types of DNA damage have been studied as well by US and other enhanced sampling techniques. The base flipping of undamaged DNA bases in DNA alone and in complex with cytosine-C5-methyltransferase from HhaI has been simulated by

Umbrella Sampling simulations [71] . These simulations clearly indicate that a trapping of spontaneous flipped bases [141, 150, 180] is less likely in comparison to a mechanism involving active facilitation by Modification methylase HgaI (M. HgaI) and thereby reducing the free energy barrier of flipping through the major groove. This simulation also showed a stabilization of the fully flipped state [71]. In DNA lesions generated by polycyclic aromatic hydrocarbons, the recognition mechanism involves the β-hairpin of XPC-hHR23B nucleotide excision repair and is accompanied by partner base flipping. This has been studied in the case of the Saccharomyces cerevisiae (S. cerevisae) homologue by Umbrella Sampling simulations of partner base flipping of the lesion. In comparison to undamaged DNA, the free energy barrier of flipping is in the order of $10\,kcal/mol$ in comparison to $18\,kcal/mol$ for undamaged DNA [186].

The effect of the damage on the bending angle of DNA has been studied in particular. Here, simulations have indicated that lesions can reduce the bending force constant and can thereby be used by repair enzymes to induce DNA bending and facilitate lesion flipping [130] Adaptive Umbrella Sampling simulations using the roll angle to study the bending of DNA double-helices have shown linear dependence on larger scales of bending which suggests an increase in flexibility at bending values [155].

An overview about these results can be seen in Table 4.1. Finally, it can be said that many prior experiments and simulations have shown promising results in the attempt of explaining the repair and recognition mechanism of CPD damages in DNA. However, many details of the processes are still unknown and were studied during the course of this thesis.

Part II

UNRESTRAINED SIMULATIONS

The following two chapters will presents results obtained from unrestrained simulations. Chapter 5 shows that spontaneous flipping from the extra- to the intra-helical state can be observed in free MD simulations. The backbone configurations of damaged DNA show a large difference in comparison to B-DNA. Chapter 6 has been published in similar form in the journal Biopolymers in 2014 [89]. Herein, extensive comparative molecular dynamics (MD) simulations on duplex DNA with central regular or CPD damaged nucleotides were performed. In contrast to the first simulation of Chapter 5, no flipping could be observed for the simulations which started in the intra-helical state. Hence, a clear directionality of the flipping transition is found. Further on, many helical parameters show that the isolated damaged DNA adopts transient conformations which resembled the global shape of the repair enzyme bound conformation more closely compared to regular B-DNA. Notably, it could be shown that the transient overlap of isolated DNA with the enzyme bound DNA conformation plays a decisive role for the specific rapid initial recognition by the repair enzyme prior to the looping out process of the damaged DNA. These results hint at a recognition mechanism involving the structural differences between undamaged and damaged DNA.

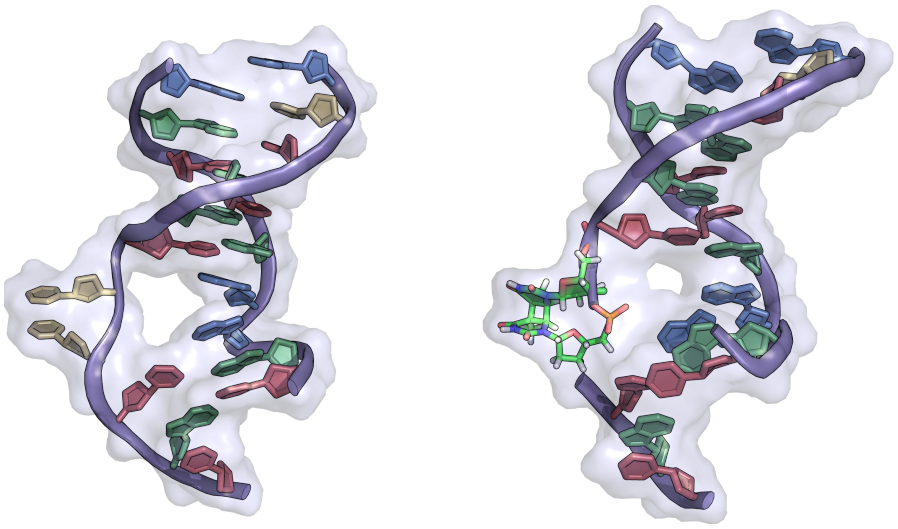
5

UNRESTRAINED MD IN ABSENCE OF REPAIR PROTEIN

5.1 INTRODUCTION

At the beginning of this thesis, multiple unbiased simulations were performed. Herein, the results from these simulations will be presented.

5.2 UNRESTRAINED MD STARTING FROM EXTRA-HELICAL STATE



(a) Starting structure of undamaged DNA. Obtained from PDB:1TEZ [117] by deleting the protein.

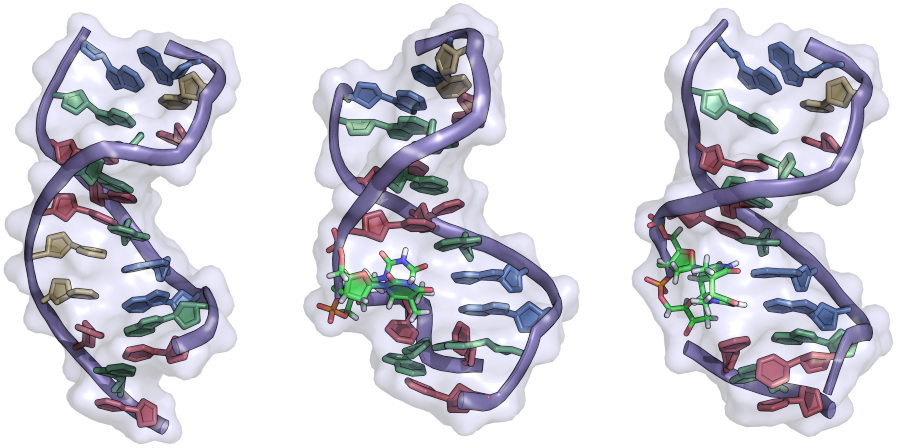
(b) Starting structure of CPD-containing, damaged DNA. Obtained from the structure Figure 5.1a by replacing the thymine bases with CPD.

Figure 5.1: The starting structures with the base in the extra-helical configuration.

As the crystal structure of PDB:1TEZ [117] (shown in Figure A.3) contains DNA in its extra-helical flipped-out state in complex with the photolyase enzyme, it was tested whether the Amber Force Field can capture the expected molecular dynamics. A flip from the externally-flipped conformation into the internally-flipped conformation (similar to B-DNA) is expected if the repair enzyme is not present. Deleting the protein resembles a situation where the protein detaches before the bases flip into the internal configuration. This obviously is only reasonable for the DNA with two thymines

as the DNA with CPD would be repaired before detaching from the protein. Nevertheless, some first interesting results can be deducted from the comparison of both simulations. Beforehand, it had been shown that the structure is stable over the time-scale of 100 ns if the protein is included. However, if the protein is not included in the simulation, the bases will flip into the internal configuration shown in Figure 5.2.

In detail, by starting from the crystal structure the protein was deleted and the structure shown in Figure 5.1a obtained. By further replacing the two flipped out thymines with CPD, the structure shown in Figure 5.1a was generated. The details are similar to the method explained in Appendix A.



(a) B-DNA configuration of undamaged DNA. All bases are in their intra-helical configuration.

(b) Intermediate configuration of damaged DNA. CPD is close to the intra-helical configuration. However, proper Watson-Crick pairing is not established.

(c) The final configuration of damaged DNA. CPD is in the best-possible intra-helical configuration.

Figure 5.2: Intra-helical conformations of undamaged and damaged DNA.

The final configuration of the undamaged structure is shown in Figure 5.2a. Before the CPD-containing DNA adopts its final structure (shown in Figure 5.2c), it flips into a different internally-flipped configuration (shown in Figure 5.2b).

To further analyze the flipping of the bases from the external configuration (Figure 5.1) into the internal configuration (Figure fig:1TEZ-internal-structures), the pseudo-dihedral angle defined in Chapter 7 (see also Figure 7.2 is analyzed Put simply, this dihedral angle measures the flipping angle of the CPD base (two thymine bases prospectively for undamaged DNA). A value close to 30° resembles a flipped-in configuration. A value of 200° represents the

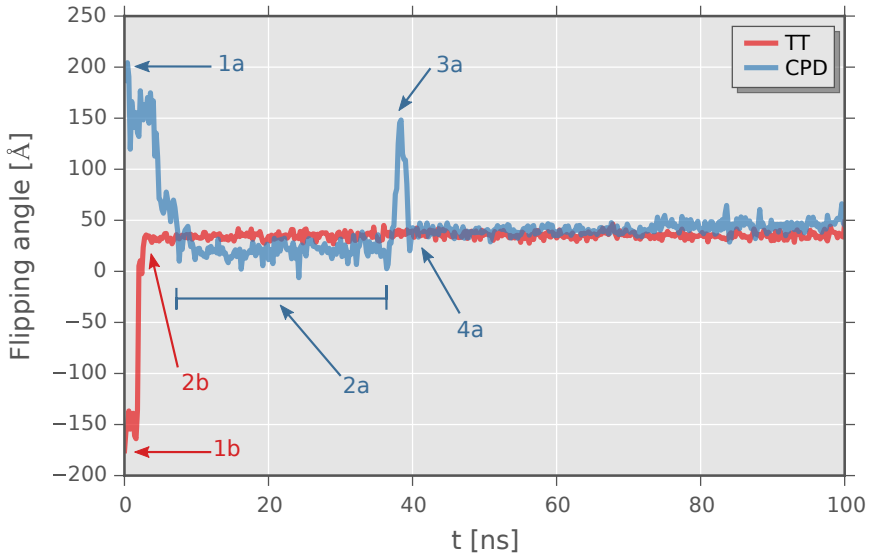


Figure 5.3: Pseudo-dihedral angle of the external bases (two thymine, CPD respectively) versus the simulation time.

externally-flipped configuration. Due to the radial symmetry of the angle, the latter value is equivalent to $-160°$ [1].

Figure 5.3 shows the pseudo-dihedral angle of the external bases (two thymine, CPD respectively) versus the simulation time. The data was plotted such that no jumps due to the 360°-symmetry appear. Multiple observations can be made. First, the undamaged DNA flips the two thymine bases in a relatively short period from the external configuration (1b in Figure 5.3, structure shown in Figure 5.1a) to the internal configuration (2b in Figure 5.3, structure shown in Figure 5.2a). The transition happens after 2 ns in the short time span of approximately 1 ns.

The dynamics of the CPD-containing DNA are more complicated. The CPD-containing DNA starts from the external configuration (1a in Figure 5.3, structure shown in Figure 5.1b). The transition to an internally-flipped structure is considerable slower compared to the simulation of undamaged DNA. After approximately 9 ns, the CPD-containing DNA is in intermediate intra-helical state (2a in Figure 5.3, structure shown in Figure 5.2b). It stays in this state for more than 15 ns before flipping back into the extra-helical configuration (3a in Figure 5.3). The damaged DNA finally adopts the intra-helical state (4a in Figure 5.3, structure shown in Figure 5.2c).

As these simulations only contain a single change in configuration, rigorous statistical analysis cannot be obtained. However, first

1 A flipping angle must not necessarily have a simple 360° symmetry. A twist of the base by flipping of 360° could also induce a different conformation in the structure as the measured pseudo-dihedral angle is a sum of multiple angles. Here, it was however checked that this is not the case and that the conformation of 360° is exactly equivalent to the conformation of 0°.

conclusions can be drawn. It could be shown that the time span for the flipping of the CPD is likely larger than the time span for the flip of two thymine bases. This is an indication that the external configuration is more stable for CPD-containing DNA. Further, also competing intra-helical configurations exist for the CPD-containing DNA. That is, the CPD dimer is already slightly pushed out into the extra-helical conformation along the major groove direction.

5.3 BACKBONE STRUCTURE OF 1TTD

McAteer et al. determined one of the first NMR structures for duplex DNA containing a CPD lesion. The setup of structure of PDB:1TTD has been explained in Section A.2. Here, the backbone conformations of a simulation of damaged DNA were analyzed and compared to its B-DNA counterpart created by the NAB tool of Amber. Unrestrained MD simulations of damaged and undamaged DNA starting from these intra-helical states were run for 100 ns with the Amber12 [27] force field using modifications by Spector et al. [153]. Snapshots of normal B-DNA and DNA containing a CPD dimer are shown in Figure 5.4.

As previously explained, the backbone of DNA is restricted in its backbone angle conformation (see Figure 1.1). In particular, the distributions of correlated backbone angle pairs are plotted. The backbone angle pairs of α-γ , ϵ-ζ, and η-θ are well defined.

However, the distributions of these pairs differs significantly between CPD-containing DNA and regular DNA. The distribution of the α-γ pair differs dramatically in the measured γ between damaged (shown in blue) and undamaged DNA (red) (Figure 5.5). Both, the pairs of ϵ-ζ (Figure 5.6) and η-θ (Figure 5.7) differ significantly in their first angle (ϵ, respectively η) between damaged and undamaged DNA.

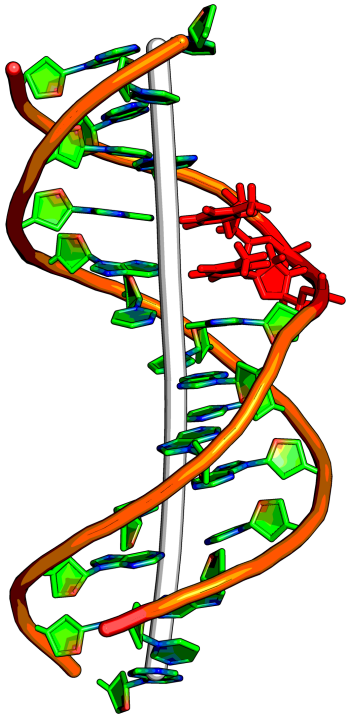
Overall, this results in significant DNA bending (27°) which is on average in good agreement with experimental values of 30° [113]. In comparison, B-DNA is bend by about 9° in the Amber force field. The bending can be easily observed in the snapshots (Figure 5.4). Further a change of the width of the minor groove is observed. Among other things, the details of the minor groove are analyzed in Chapter 6.

5.4 CONCLUSIONS

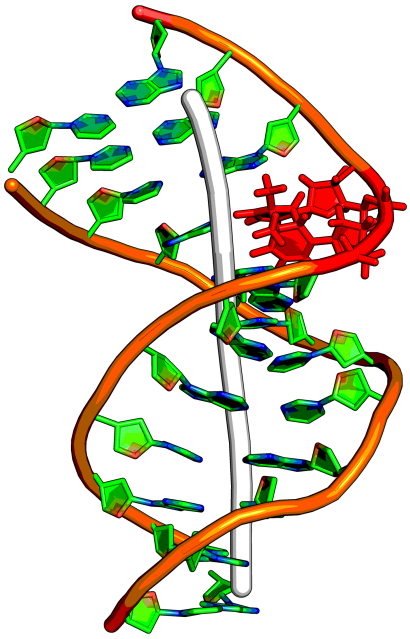
In none of the simulations a spontaneous flip of the CPD lesion (or undamaged thymine bases) from the properly paired intra-helical configuration into the extra-helical configuration could be observed. Together with the observed flipping from the extra- to the intra-helical state it may be deduced that the extra-helical state is of higher

free energy than the intra-helical state. This preference is lower for the damaged DNA, making the transition to extra-helical state faster.

Further on, significant differences in backbone configurations between damaged and undamaged DNA have been observed and may be responsible for the differences in free energy. A more detailed analysis of much longer simulations of the intra-helical starting structure will follow in the Chapter 6.



(a) B-DNA is only bend to a small amount.



(b) DNA containing a CPD lesion is bend by 27°.

Figure 5.4: Difference in bending by the introduction of a CPD lesion into double stranded DNA.

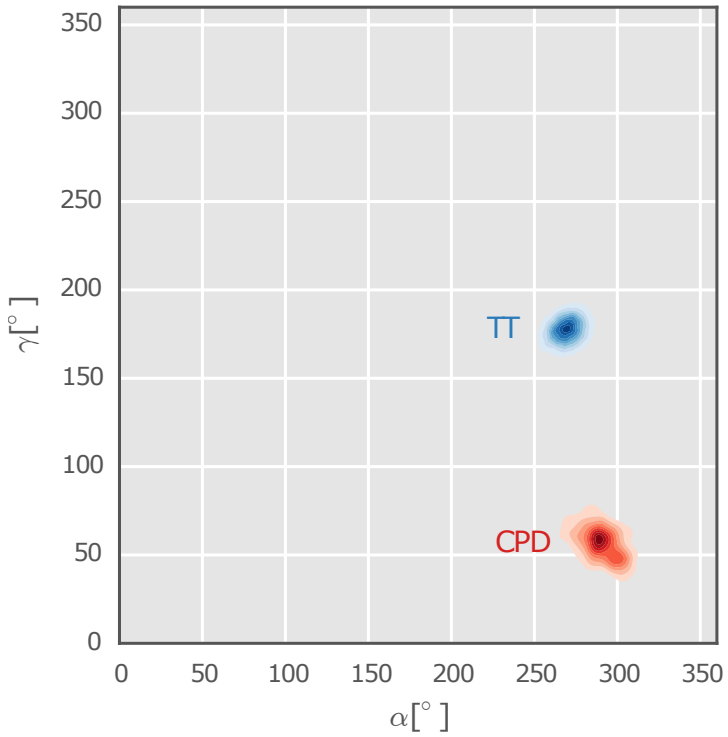


Figure 5.5: Distribution of α-γ backbone angle pair of damaged and undamaged DNA.

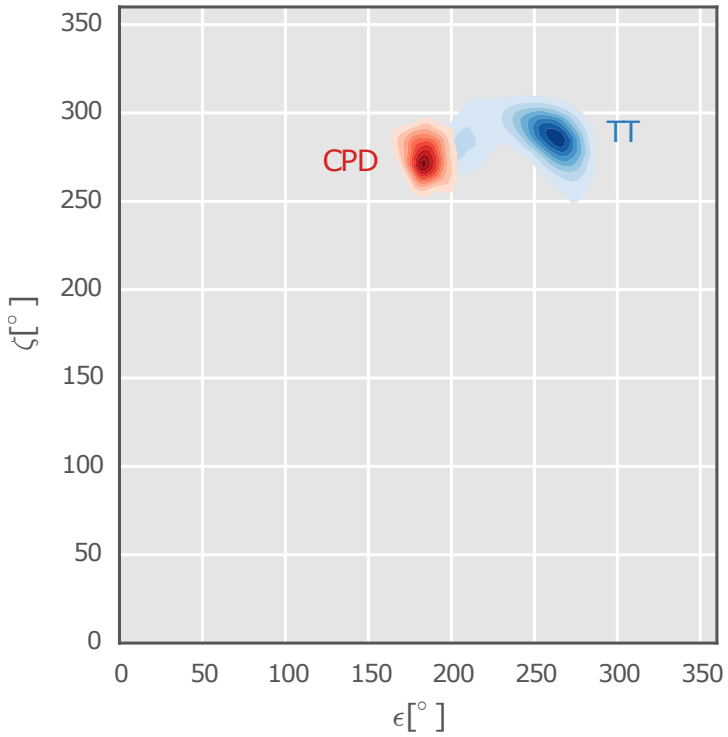


Figure 5.6: Distribution of ϵ-ζ backbone angle pair of damaged and undamaged DNA.

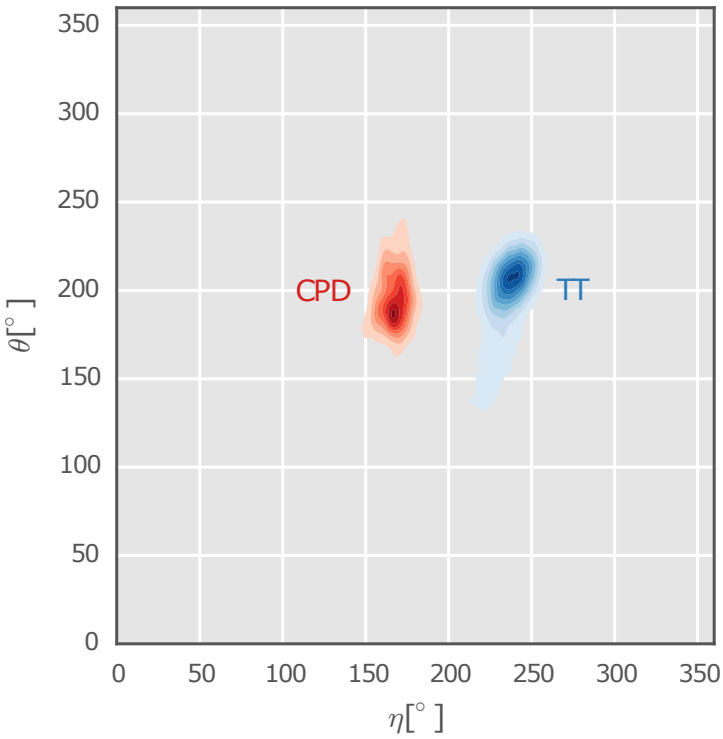


Figure 5.7: Distribution of η-θ backbone angle pair of damaged and undamaged DNA.

6

EXTENSIVE MD STARTING FROM INTRA-HELICAL STATE

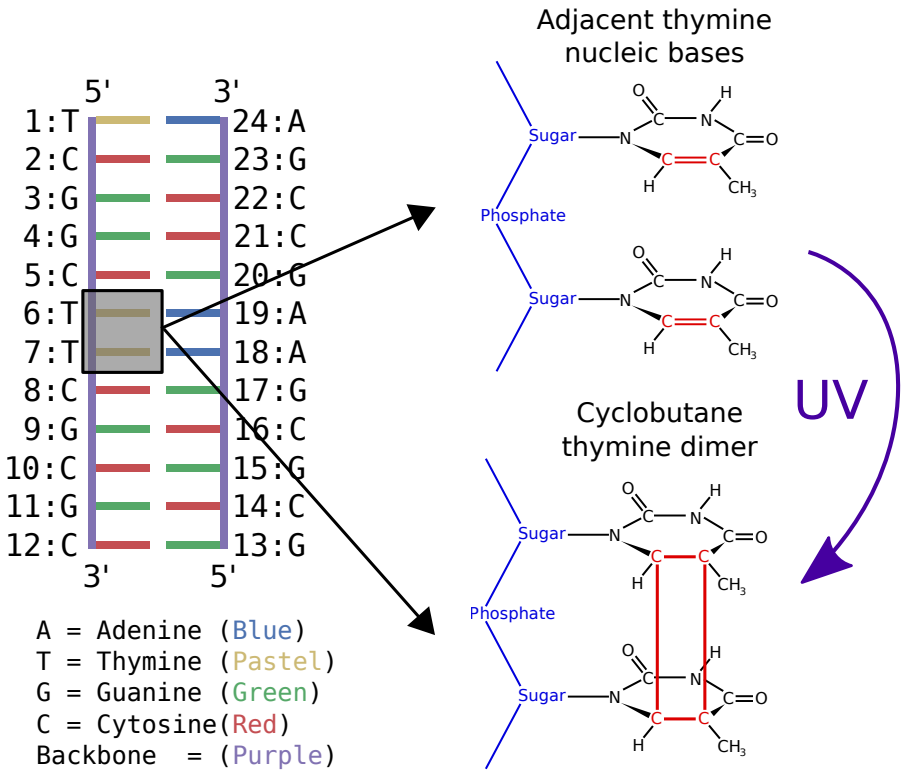
6.1 INTRODUCTION

Important observations about the repair mechanism of CPD damaged DNA can be obtained by studying its local structure. In particular, the effect of the damage on the change in structure is of special interest. The crystal structure of DNA containing CPD lesions in complex with DNA photolyases has been determined and indicates a significant distortion of the DNA. The damaged DNA is recognized through binding to an enlarged minor groove. The DNA is strongly bend and under-wound and the damaged CPD bases are flipped out into an extra-helical conformation to enter the enzyme active site [117].

As explained in Section 1.3, the extra-helical state of the damaged site is required for the enzymatic repair as the electron pathway requires close contact of the excited Flavin Adenine Dinucleotide free radical (FADH) transfers to the pyrimidine dimer [146].

However, a recognition mechanism (see Chapter 3 and especially Figure 3.3) which relies on structural changes induced by the CPD lesion, is possible.

Indeed, both, X-ray crystallography as well as NMR spectroscopy indicate a significant damage-induced distortion of the helical DNA structure with respect to regular B-DNA in the absence of a bound repair enzyme (PDB:1N4E) [133]. This has been additionally confirmed by gel electrophoresis and electron microscopy which have provided evidence for a partially distorted and bend structure of isolated DNA which contains CPD lesions[75, 173, 174]. The degree of DNA bending and DNA unwinding varies in different experimental studies indicating that the CPD damage creates probably a range of possible conformational states in DNA. In this process binding of the repair enzyme onto DNA lowers the free energy barrier of the flipping process. It is supported by the fact that all currently available structures (including solution NMR structures) indicate an intra-helical conformation of the CPD damage in the absence of a repair protein. This hypothesis would be backed by CPD-containing DNA which adopts transient conformations that fit into the binding site on the repair enzyme while keeping the intra-helical conformation of the dimer lesion. Such characteristics would allow an initial preferential encounter recognition following mechanistically a conformational selection process.



(a) Base pair steps and sequence of the DNA oligonucleotides. The duplex molecules included either two standard T or a central CPD damage (indicated by an enclosing box) opposite to the two adenine bases.

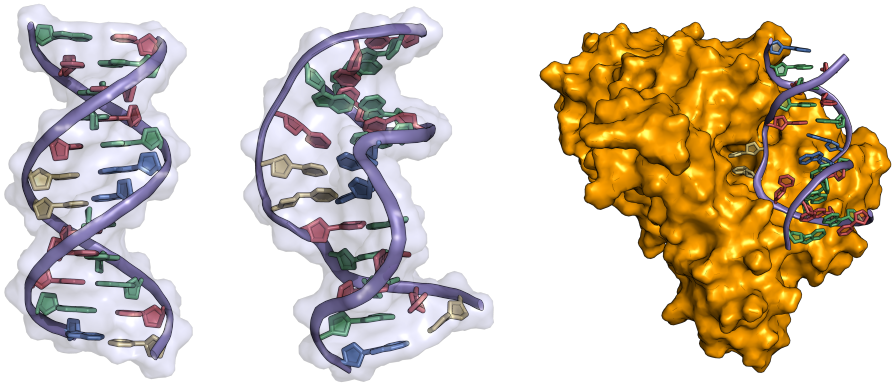
(b) Chemical structure of two adjacent thymine bases. By irradiation with UV light, thymine dimer are formed. For reasons of clarity, the newly formed dimer bonds are not in proportion to the chemical structure. In reality, the bases will change their orientation.

Figure 6.1: Sequence of the used DNA structure and details of CPD damage.

It would also explain a recognition mechanism largely based on the damage structure rather than detailed sequence of the DNA. The subsequent looping out transition of the CPD into the enzyme active site can then be considered an induced fit step of the repair process. For other repair processes, e.g. the recognition of oxidatively damaged guanine to 8-oxo-guanine, such encounter complexes with strongly bend DNA and importantly an intra-helical damaged *8oxoG* have indeed been structurally characterized [23].

In order to better understand the recognition process and the extraordinary specificity of DNA photolyases for CPD damages (Figure 6.1b) compared to regular DNA (Figure 6.2a), it is important to investigate the dynamics and possible conformations of CPD-damaged DNA in the absence of a bound enzyme. The sequence of the studied DNA is shown in Figure 6.1a. In damaged DNA, the

two thymine bases are replaced by a cyclobutane thymine dimer as shown in Figure 6.1b.



(a) B-DNA form of the DNA. Some simulations were started from this structure. The undamaged structure is labeled as $\mathbf{TT_{BDNA}}$ and the damaged structure as $\mathbf{CPD_{BDNA}}$.

(b) Deformed DNA structure by restraining all backbone atoms (except the damaged nucleotides) to the CPD-damaged DNA crystal structure in complex with the E. coli photolyase (pdb-entry: PDB:1TEZ [117]) The undamaged structure is labeled as $\mathbf{TT_{1TEZ}}$ and the damaged structure as $\mathbf{CPD_{1TEZ}}$.

(c) Slightly modified structure of PDB:1TEZ [117]. The DNA is in a complex with the photolyase protein (orange surface representation).

Figure 6.2: Structures used for MD simulations.

Comparative explicit solvent MD simulations of up to 1000 ns on regular undamaged B-DNA and DNA containing a central CPD damage were employed to elucidate the dynamics and the range of possible conformational states. In order to evaluate the overlap with conformations in the presence of a bound DNA photolyase, simulations were also performed on the DNA in complex with the repair enzyme and the central nucleotides flipped out into the enzyme active site (structure of flipped out state shown in Figure 6.2c). Regular undamaged DNA was found to adopt conformations in the absence of the repair enzyme that deviated significantly from the global conformation of the damaged DNA in the bound state. However, for the isolated CPD-containing DNA some overlap of sampled states with states in the enzyme bound form were observed. The results indicate that the initial recognition likely follows a mixed conformational selection and induced fit process.

6.2 METHOD

The system was set up as explained in Appendix A. The resulting starting structures are shown for the undamaged case in Figure 6.2. The simulations were started from four different starting configurations. The B-DNA like structures are termed $\mathbf{TT_{BDNA}}$ and $\mathbf{CPD_{BDNA}}$, respectively ($\mathbf{TT_{B\text{-}DNA}}$ shown in Figure 6.2a). DNA structures in the intra-helical protein-binding conformation are termed $\mathbf{TT/CPD_{1TEZ}}$ ($\mathbf{TT_{1TEZ}}$ shown in Figure 6.2b).

In both cases, the starting structures contained the central damaged or undamaged bases in an intra-helical state (see Appendix A for more details) resulting in four separate simulations of the DNA oligonucleotides in the absence of the repair enzyme.

Analysis of helical parameters was performed with Curves+ [95], python, *scipy* [78], and *numpy* [168]. The graphs were plotted with the help of the *matplotlib* library [74].

6.3 RESULTS AND DISCUSSION

Molecular dynamics simulations starting from B-DNA or repair enzyme bound structures

Molecular dynamics (MD) simulations of CPD-damaged and undamaged DNA of the same sequence were performed of a simulation length of 1000 ns in the absence of the DNA photolyase repair enzyme.

The simulations starting from different initial structures (with an initial root mean square deviation of the backbone RMSD of 5 Å for the central six base pairs) converged to a similar final distribution of conformations with the same RMSD distribution with respect to B-DNA (Figure 6.3). The same result was obtained if the RMSD evolution was compared to the structure found in the crystal structure of an unbound CPD-containing DNA (PDB:1N4E [133]) or to the backbone structure in complex with the photolyase repair enzyme (Figure 6.3). Interestingly, the relaxation to similar RMSD levels was much faster in case of the regular undamaged DNA (few ns) compared to the DNA with the central CPD damage ($\approx$ 10 ns–20 ns). Although the simulations indicated a significant flexibility of the CPD-damaged site (fluctuations in the RMSD vs. time plots, Figure 6.3) no transition to a flipped out extra-helical state was observed. As expected, the trajectories of the regular undamaged DNA approached the B-DNA reference most closely. The simulations of the CPD-containing DNA reached a final average structure closer to the CPD-containing crystal structure PDB:1N4E than to B-DNA (Figure 6.3).

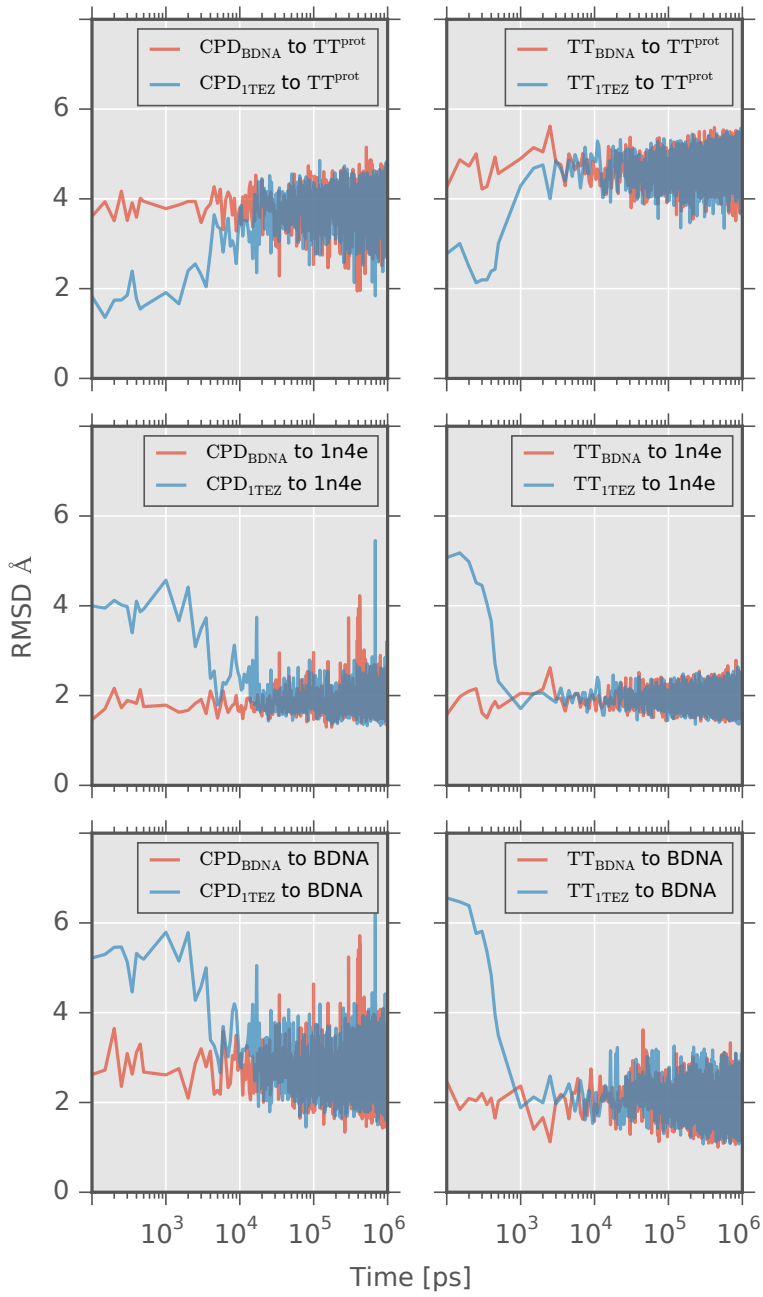


Figure 6.3: RMSD of the central part of the duplexes (backbone atoms of nucleotides 4 to 9 and 16 to 21 including the central CPD:A or T:A bases pairs) during MD simulations of CPD-damaged (left panels) and undamaged DNA (right panels). In each case the RMSD was calculated for two trajectories starting from B-DNA (blue lines) and from initial deformed DNA towards the photolyase-bound form (green line, see Methods for details). The RMSD was calculated using either the enzyme-bound DNA backbone as reference (upper panels), the crystal structure of the isolated CPD damage at the center of a double-stranded DNA (PDB:1N4E, middle panels) or using B-DNA as reference (lower panels).

For further comparison of the sampled states in the absence of the repair enzyme with enzyme bound conformations, additional MD simulations of the complex between E. coli photolyase and the DNA with the central two bases in the extra-helical state were performed for 600 ns (start structure: PDB:1TEZ, termed $\mathbf{TT}^{\mathbf{prot}}$). Here, significant differences in the distribution of the backbone RMSD (central 6 base pairs) with respect to the conformation in the photolyase bound form of CPD-damaged and regular DNA (Figure 6.4) can be seen.

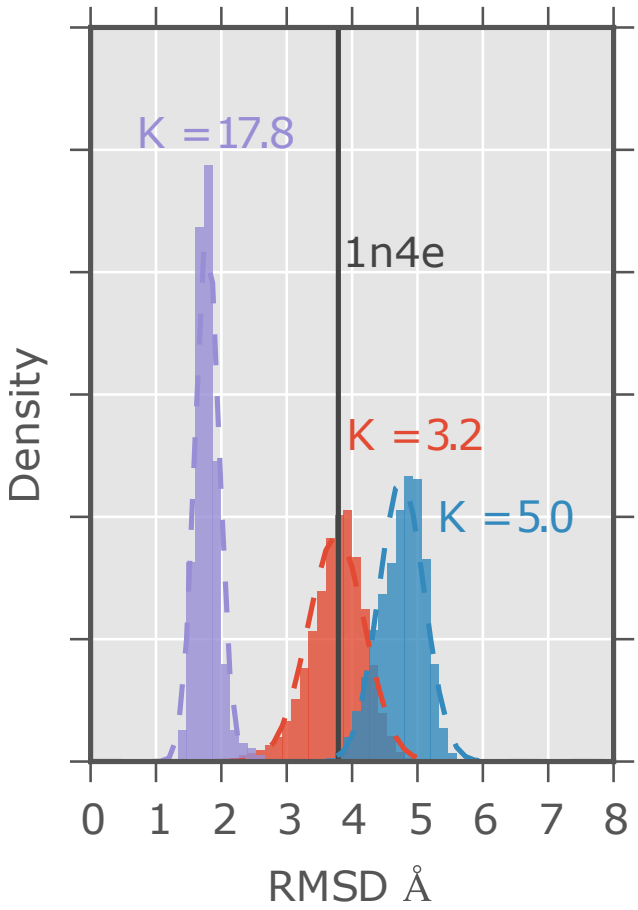
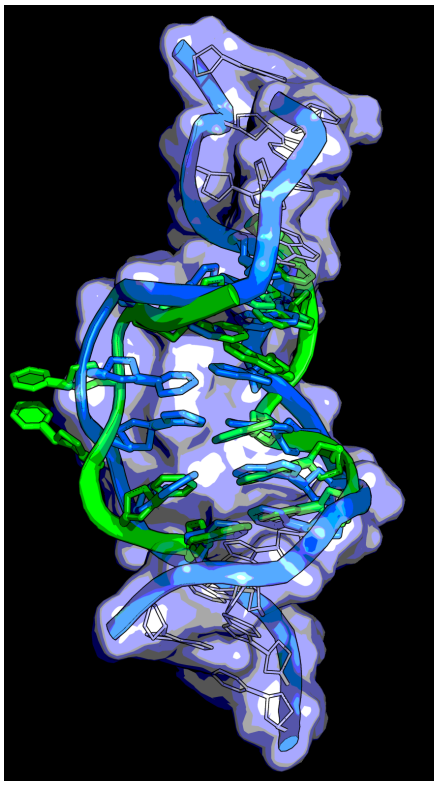


Figure 6.4: Backbone RMSD-histograms (nucleotides 4 to 9 and 16 to 21, same mask as in Figure 6.3) obtained for the MD simulations of undamaged DNA (blue histogram), CPD-containing DNA (red) and DNA in complex with photolyase (violet). The RMSD of the crystal structure of CPD-containing DNA PDB:1N4E [133] is indicated as a vertical black line. The CPD-containing DNA in complex with photolyase (bound conformation, PDB:1TEZ) is the reference. The effective force constants for deformations with respect to the enzyme bound form obtained from Gaussian fits are indicated have been used to estimate the free energy change of DNA deformation towards the enzyme bound form.

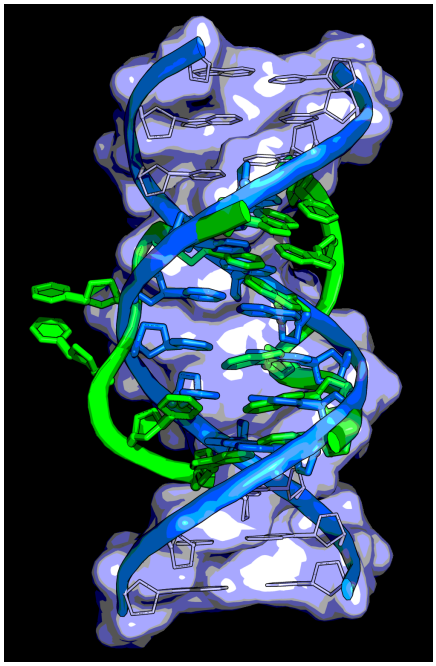
The RMSD distribution histogram obtained from the simulations of the damaged DNA overlaps to some degree with the distribution obtained for the enzyme bound form which is clearly not the case

for regular undamaged DNA (Figure 6.4). Furthermore, a direct comparison of DNA conformations sampled in the simulations of the isolated CPD damage and in the repair enzyme bound form yields a minimum RMSD of 1.8 Å compared to a minimum RMSD of 2.9 Å for the undamaged structure (Figure 6.5a), respectively. Hence, already in the absence of the repair enzyme the CPD-damaged DNA transiently adopts conformations closer to the bound form than the undamaged DNA of the same sequence.

The sampled conformation that is closest to the enzyme bound conformation was analyzed in particular. Interestingly, this conformation can be superimposed with the damaged DNA bound to the protein resulting in good overlap and little sterical clashes with the photolyase enzyme molecule (Figure 6.5b).



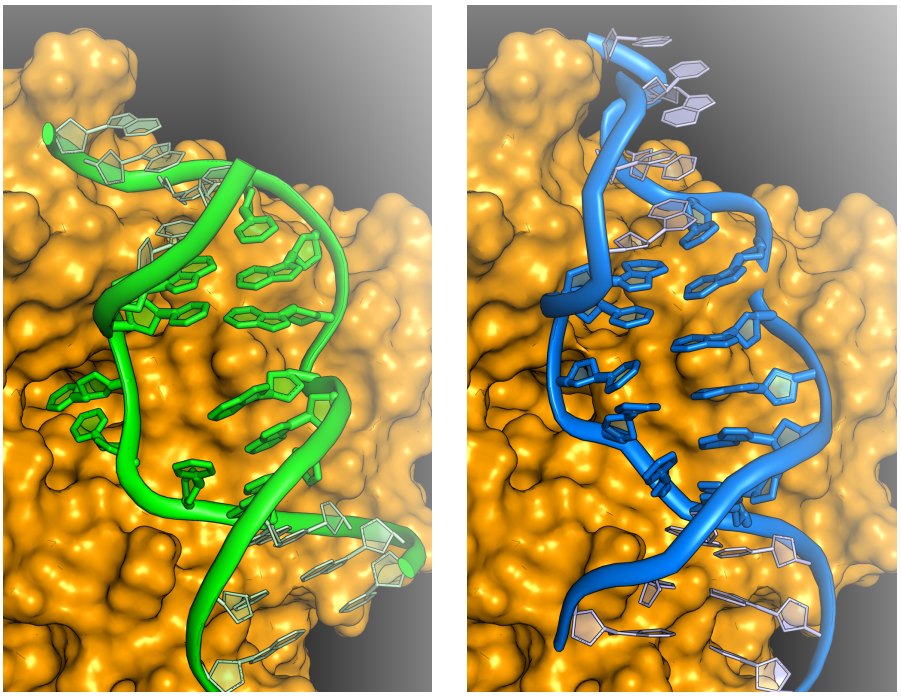
(a) Snapshot of the simulation of the CPD-damaged DNA (blue) which approached the enzyme bound conformation (green, only the six central base pairs are shown) closely with an RMSD (backbone of the central six base pairs) of 1.8 Å after best superposition.



(b) Superposition of regular B-DNA (blue) onto enzyme bound damaged DNA, otherwise same as Figure 6.5a.

Figure 6.5: Sequence of the used DNA structure and details of CPD damage.

These results reinforce the view of a possible structure based recognition process. The following analysis was done in reference



(a) Close up view of the X-ray structure of the CPD-damaged DNA (green) in complex with photolyase (orange surface).

(b) DNA conformation snapshot described in Figure 6.5a after superposition onto the DNA in complex with photolyase (only the superimposed DNA structure (blue) and the photolyase enzyme (orange surface) are indicated.

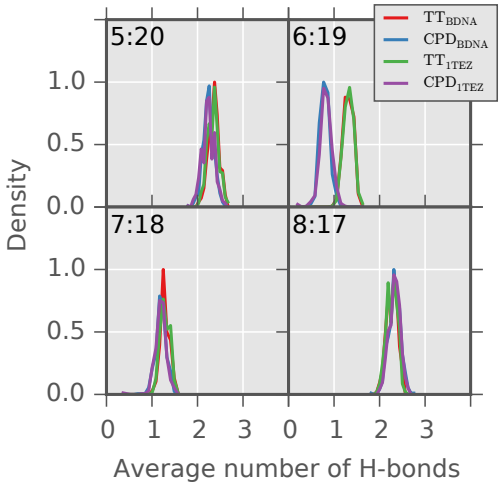
Figure 6.6: H-bonds and responsible configuration.

to the CPD-containing DNA as shown in Figure 6.4. The observed RMSD distribution histograms for the sampled CPD-containing DNA and the regular DNA from the enzyme bound form can be fit well by Gaussian distributions (Figure 6.4) suggesting a quadratic dependence for deviations from a mean RMSD. Assuming that such quadratic dependence is also valid up to deformations that come close to the enzyme bound form, a rough estimate of the energetic contribution for such induced fit of the DNA upon initial binding by the repair enzyme can be calculated. As a target RMSD $R_{target} = 1.8\,\text{Å}$ is used to represent the mean of the protein bound conformation (Figure 6.4). With these assumptions the calculated deformation penalty for the damaged DNA is much smaller ($\approx$ $6.4\,kcal/mol$) compared to the undamaged regular DNA with central T:A base pairs ($\approx$$22.5\,kcal/mol$). It is likely that these penalties are overestimated by a simple quadratic model, however, the large difference between damaged and undamaged DNA indicates that the difference in conformational distribution between damaged and

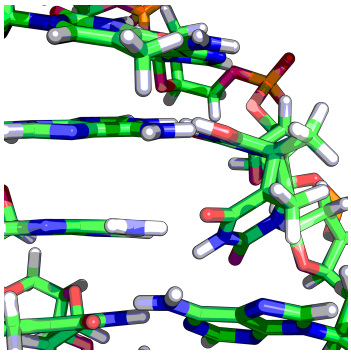
undamaged DNA already in the absence of the repair enzyme plays a significant role for the recognition process and binding affinity.

Nucleo base hydrogen bonding at the damaged site

The average number of intra-helical hydrogen bonds formed by the damaged bases is one important factor that determines the stability of the intra-helical CPD-damaged structure. Crystal structures and NMR structures of CPD-damaged DNA indicate a reduction of the number of hydrogen bonds of the CPD bases with the adenine bases on the opposite strand compared to regular T:A base pairs. Indeed, during the simulations of the undamaged DNA around one hydrogen bond per base pair was observed and the distribution was similar for both central T:A base pairs (Figure 6.7a).



(a) Distribution of intra-base-pair hydrogen-bonds of the central two base pair steps observed during MD simulations of CPD-damaged and undamaged DNA starting from different starting structures (gray and blue lines, respectively). Numbers $n : m$ at the top left define the H-bonds between base n and m.



(b) Asymmetry of hydrogen bonding geometry in the crystal structure PDB:1N4E [133] indicating the central CPD damage and directly flanking nucleotides (atom color coded stick model).

Figure 6.7: Hydrogen bonding patterns.

In case of the CPD-damaged DNA for the 3-A:CPD base pair only a slight reduction of the average number of H-bonds was observed. However, for the 5'-base pair a significant reduction to, on average, 0.75 H-bonds was found. Therefore the 5'-base pair $6 : 19$ is affected more by the CPD lesion. The asymmetry of the hydrogen bonds is related to the backbone having a distinct direction and agrees with the crystal structure of isolated CPD-damaged DNA PDB:1N4E [133] (Figure 6.7b). Interestingly, the H-bond distribution of the directly

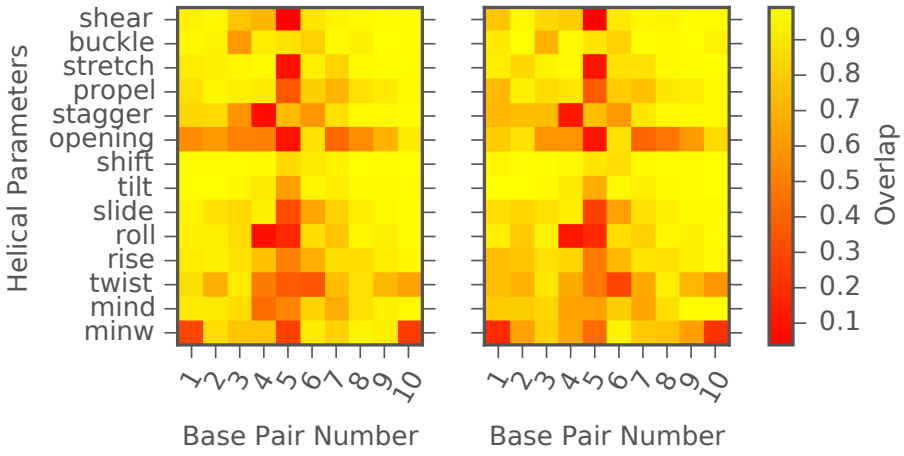
flanking G:C base pairs is not affected by the absence or presence of the central CPD-damage.

Analysis of helical geometry at the damaged site

It is convenient to describe the structure of double stranded DNA in terms of helical parameters associated with each base pair (6 intra base pair parameters: buckle, propeller, opening, shear, stretch, stagger) or base pair step (6 inter base pair parameters: tilt, roll, twist, shift, slide, rise) assuming rigid nucleo-base conformations (for details refer to Section 2.13.3). For the recognition of damaged DNA, differences with respect to undamaged DNA are of particular interest. An appropriate measure for such differences is the normalized overlap of the helical parameter distributions sampled during MD simulations of damaged and undamaged DNA (Figure 6.8). Very similar overlaps between helical parameter distributions were obtained when comparing CPD-damaged and undamaged DNA starting from the two different start structures (Figure 6.8) indicating good convergence of the sampled states. As expected, there are significant differences between some of the helical parameters of the distributions of damaged and undamaged DNA (little overlap between distributions, Figure 6.8). However, already for the nearest neighbor base pairs and next-nearest neighbors the overlap of most helical parameter distributions was almost 1, indicating that the helical deformation due to the damage is mostly localized directly to the central damaged site and affects neighboring base pairs already to a much smaller degree. Large differences between distributions for CPD-damaged and undamaged DNA at the damage site were found for several helical parameters including base pair opening, roll, shear stagger, stretch, and twist.

To further investigate these differences, the base pair parameter distributions averaged over the central four base pairs are illustrated in Figure 6.9. These parameters determine the global helical geometry at and around the damaged site that is recognized by the photolyase. Again, comparison of simulations starting from different initial conditions resulted in very similar distributions of the corresponding helical parameters (Figure 6.9). Interestingly, the width of the distributions are similar for CPD-containing DNA and for regular DNA indicating that the damaged site adopts a different average structure but the magnitude of fluctuations is similar to regular B-DNA.

For the recognition it is of importance how closely the helical parameter distributions of the damaged DNA in the absence of repair enzyme resemble the enzyme bound geometry. Especially, the parameters roll, twist, and slide resulted in a distribution significantly shifted with respect to regular undamaged DNA and more closely



(a) Overlap of helical parameters calculated for simulations starting from standard B-DNA.

(b) Overlap of helical parameters for simulations starting from the near-bound form of the DNA close to the structure in complex with photolyase PDB:1TEZ [117].

Figure 6.8: Overlap of calculated helical parameter distributions at the six central base pairs (or base pair steps in case of inter base pair step parameters) of damaged and undamaged DNA simulations. The overlap was normalized such that a value of 1 (yellow) indicates a perfect agreement between distributions and 0 (red) indicates no overlap.

approaching the distribution observed for damaged DNA in complex with the photolyase enzyme. No significant shift for CPD-damaged DNA in the distributions of shift, rise, and tilt with respect to regular DNA was observed. Note, however, that the global structure of DNA is less sensitive to small changes in rise or shift compared to a roll or twist change of a DNA segment.

The distributions of the intra-helical parameters show changes of smaller extent (Figure 6.11).

The binding of a damaged DNA involves the DNA minor groove and the depth and width parameter distributions observed for the MD simulations of damaged DNA show a shift towards the distributions obtained for the simulations in complex with the enzyme (Figure 6.10). Apparently, further minor groove deformation induced by the repair enzyme during binding is still required (see also Figure 6.10).

6.4 CONCLUSIONS

The present comparative simulations indicate significant differences in the conformational states of regular B-DNA and DNA with a central intra-helical CPD damage opposite to adenine nucleotides. Already in the absence of a repair enzyme the damaged DNA can transiently adopt conformations resembling the enzyme bound form

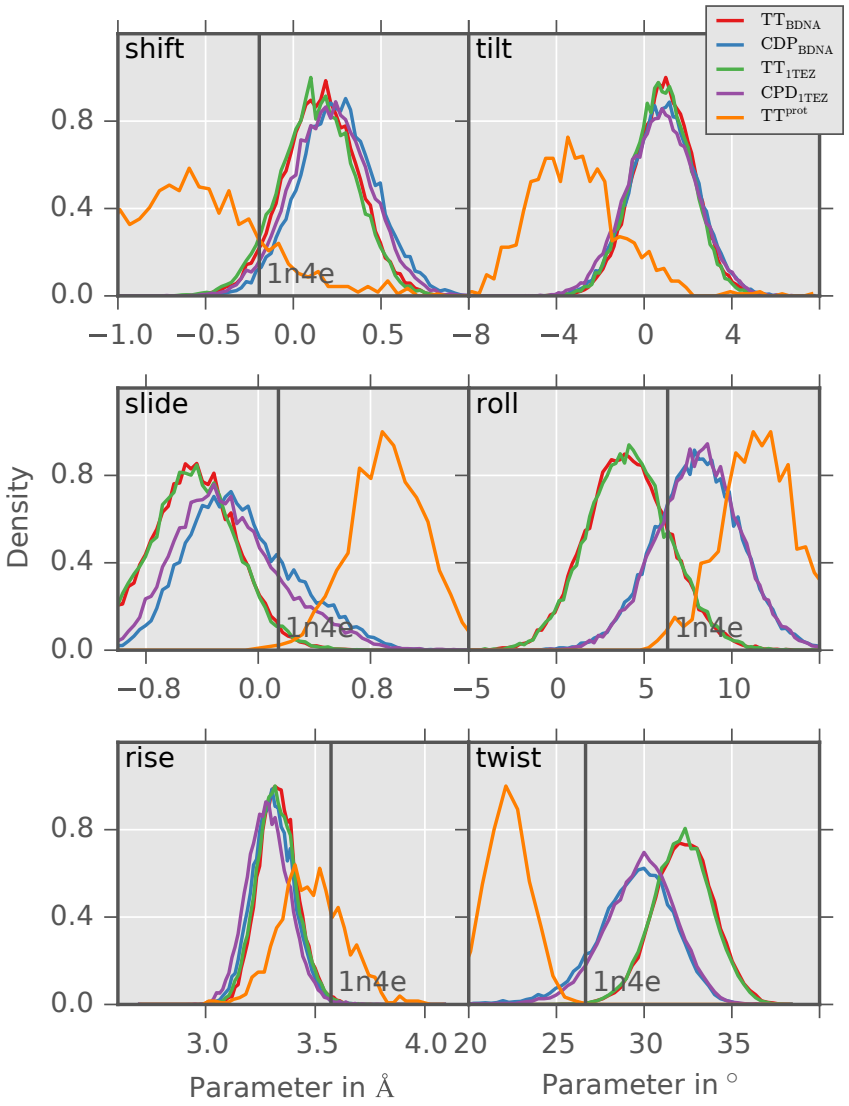


Figure 6.9: Comparison of average inter-helical parameter distributions of the central four DNA base pairs (base pairs 5-9) for simulations of damaged, undamaged and enzyme bound DNA. The average parameters give an overview on the overall helical structure of the segment that is recognized by the repair enzyme. A comparison to the distribution obtained from simulations of the complex with photolyase is also included (orange line). The helical parameters of the CPD-damaged DNA crystal structure (PDB:1N4E) is indicated as vertical line.

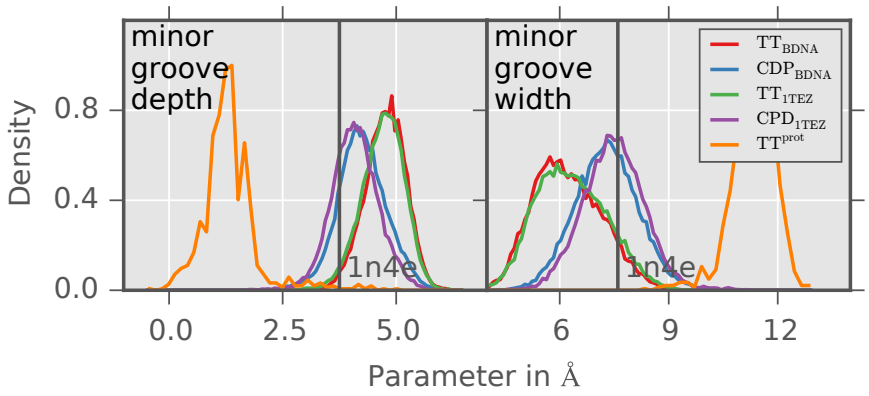


Figure 6.10: Comparison of the minor groove width and depth for the simulations of damaged, undamaged and enzyme bound DNA.

much more closely than regular DNA. This can dramatically enhance the possibility for transient encounter binding of such sites by a repair enzyme that largely recognizes the overall shape of the damaged site. However, no spontaneous looping out of the damaged bases was observed and also no increased opening fluctuations of the damaged bases. This result supports a two step recognition process with an initial binding of CPD-damaged DNA by a photolyase repair enzyme in an encounter complex in which the enzyme recognizes the global shape of the damaged site but with the damaged bases still in an inter-helical state. Such DNA conformations approaching the global shape in the enzyme bound form were sampled during the MD simulations of the isolated damaged DNA. In a second step of the repair process the enzyme promotes the transition of the damaged bases into an extra-helical conformation bound to the enzyme active site. Whether the enzyme indeed facilitates such looping out process will be further investigated by studies in which the looping out process is studied using enhanced MD methods.

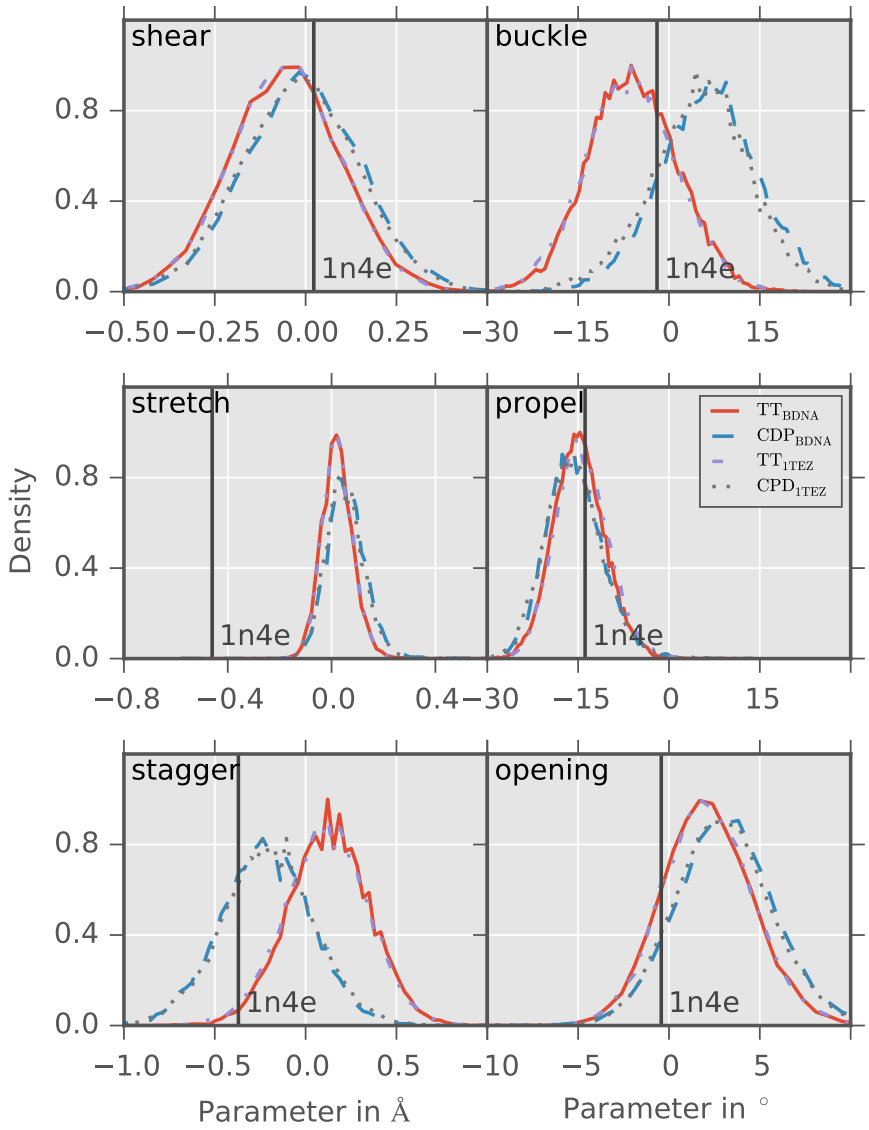


Figure 6.11: Comparison of intra-helical helical parameters of damaged and undamaged DNA. The same bases as in Figure 6.9 were studied.

Part III

UMBRELLA SAMPLING SIMULATIONS OF FLIPPING

For all the previously presented models, it is of particular interest to know the free energy profile of the flipping process from the intra-helical state of the considered bases to the extra-helical state. For the passive recognition and repair mechanism, the study of flipping mechanism in the absence of the protein will tell if this process is likely and is an effective repair mechanism. The same simulation done in the presence of the protein gives insight to the active repair mechanism. Here, the protein lowers the barrier of the flipping process of damaged bases without altering it for undamaged bases. In the following, the presented models will be rigorously tested using Umbrella Sampling MD simulations. This chapter is submitted in different and shorter form for publication.

7

UMBRELLA SAMPLING SIMULATIONS OF FLIPPING

7.1 INTRODUCTION

As shown in previous chapters, MD simulations comparing regular B-DNA and a CPD lesion in the same sequence context indicate distortions caused by the CPD damage and increased DNA flexibility. During these simulations the damage remained intra-helical without any spontaneous flipping to a looped out state during approximately 1 µs simulations time [89]. Similar results have been obtained by Masson and Laino and Miaskiewicz et al. [112, 119].

These results are alone not decisive which of the presented recognition mechanism of Chapter 3 describes the recognition of CPD damages in E. coli by photolyases best. More of the sub-processes shown in Section 3.1 have to be simulated. In particular, it is of most interest to know what the free energy difference of the intra- and the extra-helical conformation of damaged and undamaged DNA is. The quantitative predictions given in Section 3.2.2, Section 3.2.5, and Section 3.2.6 can be tested with free energy methods (see Section 2.12 for details). If an appropriate reaction coordinate is chosen to describe the transition from the intra helical to the extra helical conformation of a CPD lesion during MD simulation, the free energy changes during this transformation can be calculated. Several such free energy simulations on regular and chemically modified nucleobases have been performed indicating free energy penalties between $8\,kcal/mol$–$15\,kcal/mol$. For the flipping of the cis,syn-CPD inside a duplex DNA [128, 129, 152] a free energy change of approximately $6\,kcal/mol$–$7.5\,kcal/mol$ was obtained [129]. For the flipping of a methylated base in a complex of DNA with a DNA methyltransferase a lowering of the free energy penalty was observed [71, 170] and in human DNA by NER using the Xeroderma Pigmentosum Complementation group yeast Rad23 complex (XPC-RAD23B) protein [24, 186]. Similar, for the 8-oxo Guanine (8-oxoG) free energy simulations in the absence and presence of a repair enzyme indicated that the presence of the Formamidopyrimidine DNA glycosylase (MutM) repair enzyme facilitates the looping out of a damaged (and undamaged) base [23][1].

Here, additional Umbrella Sampling simulations were performed to study the influence of the repair enzyme onto the conformational transition of the lesion towards an extra helical state. This influence

1 For a more detailed explanation of these studies refer to **??**.

is twofold. Binding of the enzyme can result in a deformation of the DNA (e.g. changes in bending and twisting) that may lower the barrier or penalty for a looping out process. Hence, part of the binding free energy is stored as a DNA deformation. Such a mechanism does not require any direct contacts between the damaged base and the protein. In addition, direct protein-DNA contacts may mediate or facilitate the looping out process. In order to elucidate different contributions to the recognition and base flipping process in case of a CPD damage, US free energy simulations of the flipping process of three different types of simulations were performed: 1) Unrestrained DNA, 2) DNA deformed to a structure that mimics the protein induced deformation (but without repair enzyme), and 3) DNA in complex with the photolyase repair enzyme. Comparison of the calculated free energy changes of these processes allows for explanation of the molecular mechanism and the energetic contributions of each step to the flipping transition. Control calculations on undamaged DNA help to identify damage-specific contributions.

7.2 METHODS

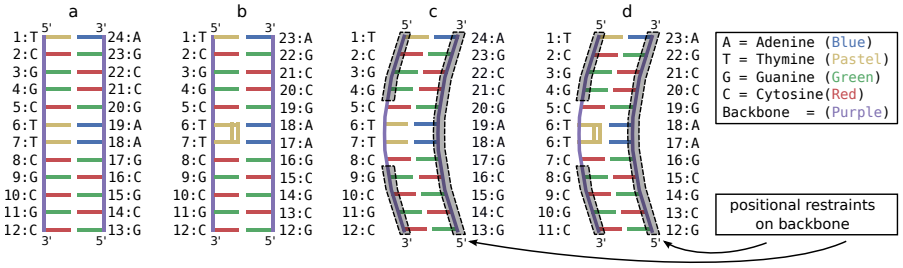


Figure 7.1: Setup: Umbrella Simulations which do not include the protein. (a) and (b) show the simulations without restraints. (c) and (d) show the simulations where positional restraints on the backbone where used to hold the DNA in the global configuration of he protein bound conformation.

In general, the same methodology as previously explained in Appendix A Section A.3 was used for the setup of the system.

MD simulations were also initiated from the complete PDB:1TEZ structure where DNA is in complex with the DNA-photolyase repair enzyme and the central bases are an extra-helical conformation bound to the enzyme active site. The exact same sequences as used for isolated DNA were used.

As a reminder, the starting structures for undamaged and damaged DNA without any restraints in the absence of the protein are denoted as $\mathbf{TT_{BDNA}}$ and $\mathbf{CPD_{BDNA}}$, the same molecules restrained to the protein bound intra-helical structure $\mathbf{TT/CPD_{1TEZ}}$, and the structures

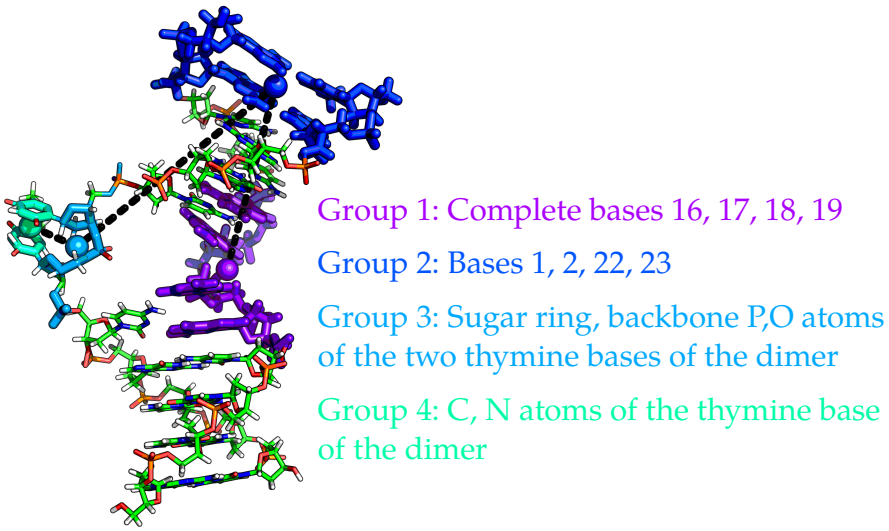


Figure 7.2: Figure showing the pseudo dihedral used for all Umbrella Sampling simulations. Participating groups were chosen to the point that mostly the damaged site is affected by the PMFs forces and not the remaining part of the DNA.

in the complex with the protein **TT/CPD$^{\mathbf{prot}}$**. For a more detailed explanation refer to Appendix A, Section A.3.

7.2.1 *Methodology of Umbrella Sampling*

In the following,the specifics of the Umbrella Sampling setup will be given. For a general introduction to Umbrella Sampling refer to Section 2.12.2. Four PMF calculations were performed in the absence of the protein. The first two started from a regular B-DNA structure for both the undamaged and the damaged DNA (see Figure 7.1 (a) and (b)). In the third (undamaged) and fourth (with central CPD damage) simulation the DNA was constrained to a bound conformation of the DNA as found in complex with the protein (denoted before as **TT/CPD$_{\mathbf{1TEZ}}$**). This was accomplished by weak positional restrains on the backbone of the bases 1 to 4 and 9 to 24 for the undamaged (c) and respectively 1 to 4 and 8 to 23 in the damaged (d) case depicted in Figure 7.1. Backbone atoms are defined as the atoms P, C5′, C4′, C3′, O5′ and O3′.

Hereby, the intra-helical configuration exhibits an angle of 29° for the **TT/CPD$_{\mathbf{BDNA}}$** conformation and 69° for the **TT/CPD$_{\mathbf{1TEZ}}$** conformation whereas the extra-helical configuration in the presence of the protein (crystal structure **TT$^{\mathbf{prot}}$**) measured an angle of −135°. Thus the angle necessary for flipping amounts to 196° (for the Thymine Thymine adjacent pair (TT)-motif respectively 156°) in the positive direction as $-135° \equiv 225°$. The positive direction represents the flipping process through the major groove. On the other hand,

Type	Group 1	Group 2	Group 3	Group 4
Damaged DNA	Complete bases 16,17,18,19	Base 1,2,22,23	Sugar ring and backbone P, O of the two thymines at damaged lesion	C, N atoms of thymine of the dimer
Native DNA	Complete bases 17,18,19,20	Base 1,2,22,23	Sugar ring and backbone P, O of two thymines	C, N-type atoms of the two thymines

Table 7.1: Definition of groups used within pseudo-dihedral angle in US.

the flipping through the minor groove requires a dihedral turn of 156° (respectively 204°).

In many systems, two important conformational states can be separated by high energy barriers. These barriers prevent the system to cross from one state to the other in the proposed simulation time. As the probability of crossing such barriers scales exponentially with negative of potential as in the Boltzmann distribution, the time to cross such a barrier might be substantially higher than the computation time. Examples include protein folding with transition times in the order of micro-seconds to seconds. Enhanced sampling method such as Umbrella Sampling [165] can mitigate this problem by introducing additional potentials to increase the sampling rate of the problematic transition states. After the biased simulation has been performed, the effect of the bias has to be reverted in order to calculate the wanted properties of the unbiased system [13, 142]. The detailed theory of this method is explained in Section 2.12.2.

Umbrella windows were simulated from an angle of $-320°$ to 38° in steps of 2° resulting in a complete 360° coverage. Figure 7.2 shows the pseudo dihedral used for all Umbrella Sampling simulations. The groups were chosen in this way to mostly increase the PMFs force on the damaged base and decrease the effect of the PMFs forces on the DNA itself. Other approaches to the selection of those groups gave miserable results where the damaged bases paired to neighbors of their actual Watson-Crick partners in the extra-helical state. In another approach with smaller groups the DNA was observed in a unnaturally distorted conformation in the extra-helical position. Using large groups for the first and second center of mass groups in the dihedral definition the latter problem was remedied. The first

problem was prevented as the axis by which the damaged bases flip is almost parallel to the axis between the second and third dihedral atom groups. The details are explained in Table 7.1.

To address the effects of specific interactions of the protein with the DNA, Umbrella Simulations in the presence of the protein were employed additionally. By addressing the overall effects of the conformation separately from the specific interactions of the protein it can be understood how the specific interaction of the protein with the DNA help in the recognition and repair process or whether the recognition happens in a different manner. The restraints used for the restrained Umbrella Sampling were the same weak positional restrains mentioned above on the backbone of the bases 1 to 4 and 9 to 24 for the undamaged (c) and respectively 1 to 4 and 8 to 23 in the damaged (d) case depicted in Figure 7.1.

In some of the production Umbrella Simulations distance restraints were employed to keep the thymine bases relatively close together to enhance sampling and to evade problems with a poor definition of the torsion dihedral whenever one base is flipped out and the other one not. The used distance restraint between the thymine nucleobases for the **TT** simulations was set to a distance in the range of 2.5 Å–6.5 Å. It is mentioned whenever this methodology needed to be used.

7.2.2 *Methodology of Hamilton Replica Exchange*

By employing HREMD the last and first windows are connected. The 360° symmetry of the systems was therefore explicitly enforced. Without the usage of this symmetry more than the 360° would have to be simulated in order to decrease boundary effects of the WHAM calculation. As pointed out in Chapter 2, Equation 2.12.5, HREMD increases sampling along space orthogonal to the chosen reaction coordinate.

By using a separate initialization procedure with a stronger dihedral restraint convergence could be improved. Initialization was run for $100\,\mathrm{ps}$ with a torsional restraint of $4000\,\mathrm{kcal/(mol\,rad^2)}$. The Production simulation was then run for $10\,\mathrm{ns}$ with a torsional restraint of $400\,\mathrm{kcal/(mol\,rad^2)}$ and 5000 exchanges and a time of $2\,\mathrm{ps}$ between exchanges. Cumulatively, the complete simulation time is $1.8\,\mu\mathrm{s}$.

7.3 RESULTS AND DISCUSSION

7.3.1 *Umbrella Sampling without the Photolyase Protein*

In order to elucidate the molecular mechanism of flipping the CPD lesion from an intra-helical to an extra-helical state, which is necessary to access the active site of a repair enzyme, umbrella sampling

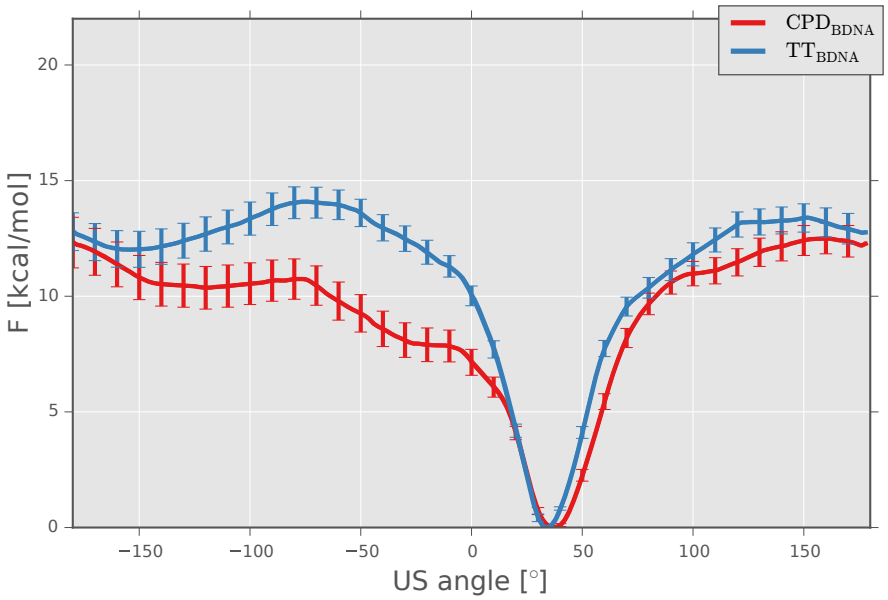


Figure 7.3: PMFs free energy for flipping process of conformation close to the native conformation.

free energy simulations along a dihedral reaction coordinate were performed to describe the flipping process (see Section 7.2.1 for details). For improving the convergence of the simulations frequent Hamiltonian replica exchanges between different umbrella sampling intervals were included (HREMD technique). A first set of HREMD simulations was performed on a double stranded DNA (ds-DNA) oligonucleotide with a central CPD thymine dimer lesion located at the center of double helix. For comparison HREMD simulations were also performed on a control ds-DNA with a flipping of two adjacent thymines and otherwise identical sequence. During the HREMD simulations no other restraints besides of the umbrella potential along the dihedral reaction coordinate were applied. The calculated Potential of mean force for the flipping process indicates an overall lower free energy penalty for flipping the CPD damage compared to a regular central TT sequence (Figure 7.3).

If one considers the larger range of looped out states relative to inter-helical states, the free energy difference between intra-helical vs. extra-helical states amounts to 9.0 kcal/mol for the TT case vs. 7.5 kcal/mol for the CPD case. This compares quite well with the experimental estimate of 9.5 kcal/mol [113] for the flipping process of CPD-damaged DNA.

The lower free energy penalty of flipping the CPD lesion is due to the fewer hydrogen bonds formed between the damaged bases and the opposite adenine residues and non-optimal stacking in the intra-helical state of the CPD lesion compared to regular DNA with central T:A base pairs. During the flipping process the two thymine bases were restraint to keep an approximately stacked conformation.

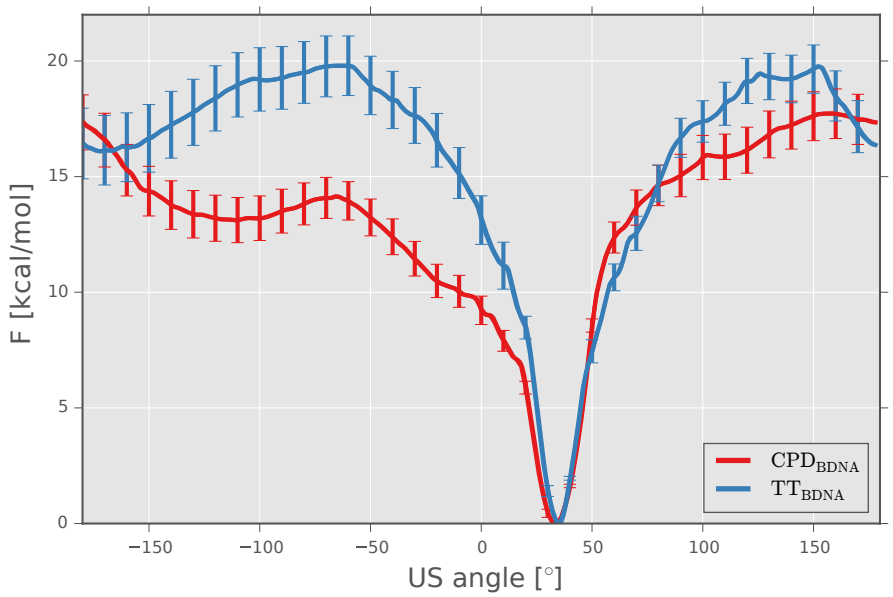


Figure 7.4: PMFs free energy for flipping process of conformation close to the B-DNA conformation. DNA is restrained to a perfect B-DNA structure.

The free energy of relaxing this restraint was calculated separately for each US window using a free energy perturbation approach (see Figure 7.3.2 and Figure 7.10). On average it amounts to less than 1 kcal/mol for the different simulation intervals (Figure 7.3.2, Figure 7.11). It is important to note that the global equilibrium structure of the CPD-containing ds-DNA differs from the regular TT containing ds-DNA. As explained in Chapter 6, CPD-containing DNA adopts structures during MD simulations that are overall closer to the structure in complex with the repair enzyme (already in the absence of the enzyme) compared to regular ds-DNA.

In order to elucidate the influence of the global DNA structure on the free energy of the flipping process, HREMD simulations were performed. These simulations included restraints to keep segments of the ds-DNA oligonucleotides in either a regular B-DNA conformer or in the conformation found in the complex with the DNA photolyase. The corresponding weak positional restraints included the nucleic acid backbone except the central lesion (or TT sequence) and the flanking nucleotides (Section 7.2.1). Specifically, the atoms P, C5′, C4′, C3′, O5′, and O3′ were used for the restraint.

In case of HREMD simulations including restraints to keep the DNA close to regular B-form the calculated PMFs were qualitatively similar to the PMFs obtained without such restraints, however, the penalty for the flipping process increased by approx. 2 kcal/mol–3 kcal/mol (see Figure 7.4). In contrast, for the simulations including restraints to deform the ds-DNA towards the global form observed in the crystal structure in complex with DNA photolyase the free energy

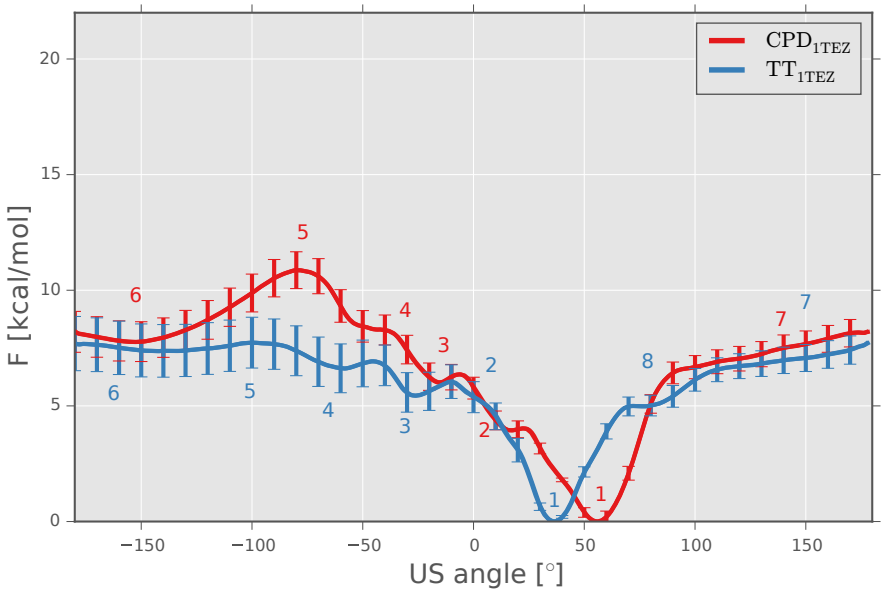
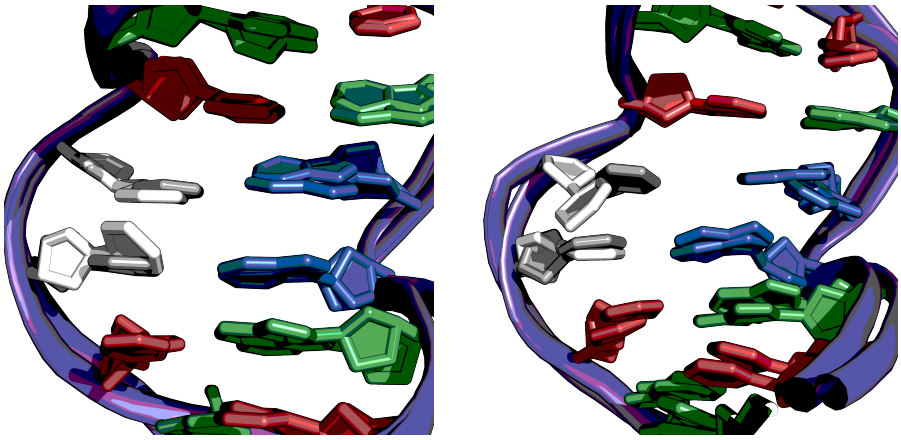


Figure 7.5: PMFs free energy for flipping process of conformation close protein bound form. For the simulation of **CPD$_{1TEZ}$**, the following steps are important: 1) WC H-bonds broken for top base. 2)Only one of the four hydrogen bonds is left. 3) All WC hydrogen-bonds are broken but configuration is still mostly intra-helical. 4) Dimer moves slowly into extra-helical configuration. It relaxes step wise 5) 90° flip into minor groove. Backbone does not hinder this flipping motion. 6) Flip of roughly 170°. 7) The movement in the major groove direction is smoother. The important steps of the US simulation of **TT$_{1TEZ}$** are the following: 1) Perfect WC pairing. 2) Breaking of bottom WC hydrogen bonds. 3) All WC hydrogen bonds are now broken for most conformations. The bases are slightly (around 45°) flipped into the minor groove. 4) Flipping into minor groove direction of 70°–90° 5) Flipping more than 90° into minor groove. 6) Completely external, flip of 180°. The backbone can relax a little bit. 7) Flipping 90° into major groove. 8) Lower WC hydrogen bonds are broken, top WC pairing is established.

penalty for flipping process decreased significantly for both the CPD and TT containing ds-DNA molecules (Figure 7.5). Interestingly, the decrease is more significant in case of the TT vs. CPD cases presumably because the CPD-containing structure adopts a conformation closer to the enzyme bound form already in the absence of the enzyme. In conclusion, the calculations indicate that a global deformation of the DNA towards the enzyme bound conformer alone (without accounting for any contacts to the protein) reduces the penalty for the flipping process significantly for both the damaged and the undamaged DNA. A restraining to regular B-DNA causes the opposite effect. Inspection of Figure 7.5 reveals another interesting effect. Whereas in unrestrained ds-DNA (or restraint to B-DNA) the free energy minimum of the intra-helical state for the CPD lesion was



(a) Snapshot of window at 38°. Both, the thymine bases (gray) and the CPD (white), are in the internal position and bound with their respective WC partners. The hydrogen bonding is better for standard thymine bases.

(b) Snapshot of window at 60°. This is the minimum free energy of the CPD flipping US simulation. In this minimum, the CPD is significantly flipped out in comparison to the thymine bases in the simulation of the undamaged DNA.

Figure 7.6: Snapshots of damaged and undamaged DNA in two intra-helical conformations.

almost identical to the position of the TT motif along the reaction coordinate (Figure 7.3, Figure 7.4), the minimum of the CPD lesion is shifted in the HREMD simulations with restraints towards the bound form (Figure 7.5).

The shift of the minimum of the internal configuration from 30° to 60° in the damaged case can be attributed to the rigidity of the damaged dimer not being able to conform to the overall bend induced by the external positional restraint. The regular DNA although under strain still keeps a hydrogen bonded geometry of the central base pairs at a 60° looping out dihedral angle (see Figure 7.6a and Figure 7.6b) whereas the CPD lesion adopts a partially flipped conformation towards the DNA major groove at the same looping dihedral angle (see Figure 7.6a and Figure 7.6b). This also creates a cavity on the minor groove side of the DNA which corresponds to the side that is bound by protein in the complex with the DNA photolyase. The attachment of the protein to this cavity will be observed in the following chapter.

7.3.2 *Umbrella Sampling in Presence of Photolyase Protein*

In addition to the HREMD simulations of isolated ds-DNA oligonucleotides simulations were also performed with the DNA bound to the DNA photolyase repair enzyme. Simulations were started from a known crystal structure of the CPD-containing DNA-repair

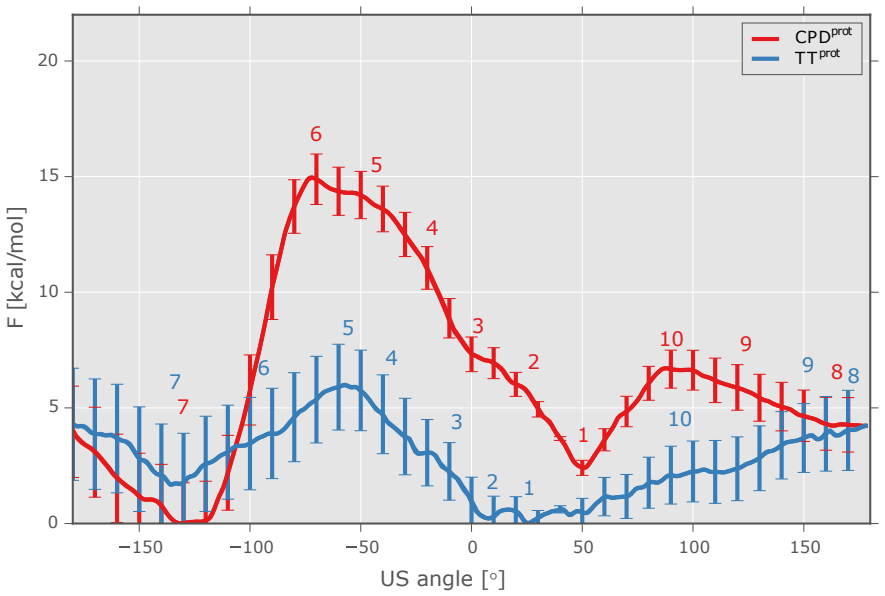
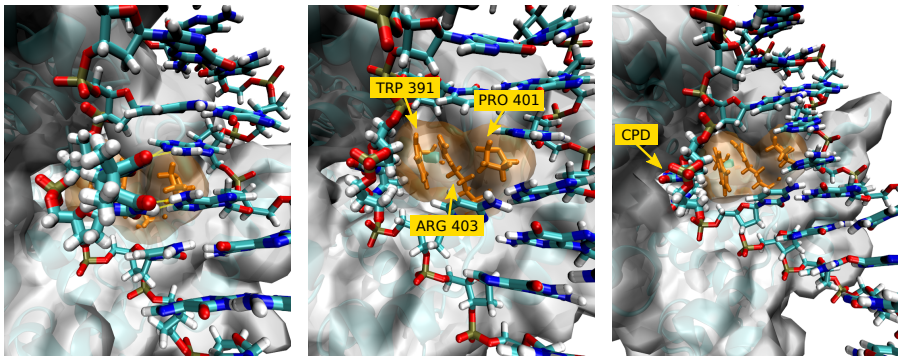


Figure 7.7: PMFs free energy for flipping process in the presence of the repair enzyme. The US simulation of **CPD^{prot}** exposes the following interesting steps: 1) This is the minimum free energy of the internal state. Note that it is already partially flipped out as also seen before in the US simulation of **CPD$_{1TEZ}$**. 2) Breaking of WC H-bonds of top base. 3) Bottom H-bonds partially broken.4) No WC pairing, slight flip. 5) Flipping started but clashes with protein. 6) Almost inside binding pocket but must overcome last clashed with protein. 7) Completely in the binding pocket. H-bonds to FADs established. 8) Flipping out (or in, depending of movement direction) of the pocket into major groove. Clashes with the ARG 254 residue of protein overcome easily as this residue is flexible. 9) 90° Flip in major groove direction. 10) WC hydrogen bonding with bottom base established. The most important steps in the flipping process of **TT^{prot}** are: 1) All WC hydrogen bonds exist. 2) Bottom WC pairing broken. Energy is low as there is apparent stacking with ARG 427 residue. 3) Clashes with PRO 425 are easily overcome. 4) Now all WC hydrogen bonds are broken. Slight flipping apparent. 5) Slowly the bases flip through the minor groove. 6) Contact to TRP 309 and TRP 425 lowering the free energy. 7) Thymine bases completely inside repair pocket. 8) As in the case of CPD, there is no big hindrance against flipping through the major groove. 9) 90° flip in major groove direction. 10) Small flip in major groove direction.

enzyme complex (including the FADHs cofactor). For the simulations with regular central TT motif the CPD lesion served as template to generate a start structure for the regular DNA case (see Section A.3). The presence of the protein significantly affected the calculated PMFs curves along the reaction coordinate (Figure 7.7). The PMFs curves indicate two free energy minima, one corresponds to the intra-helical state (at 30°) and the second minimum is located in the extra-helical



(a) CPD in the internal configuration with the maximum number of hydrogen bonds shown. PRO 401 moves in between the WC base pairs of CPD and pushes them somewhat apart.

(b) CPD has to overcome clashes with the residues TRP 391, ARG 403, and PRO 401 to move into and out (both directions) of the repair pocket. The local structure of the DNA backbone close to CPD changes in this transition.

(c) CPD in the repair pocket of the photolyase. PRO 401 is in contact with the DNA and situated in the small cavity exposed by the increased stacking distance of the WC base pair partners of the thymine dimer.

Figure 7.8: Three important configurations in the flipping process of CPD in complex with the photolyase protein.

regime and corresponds to the localization of the CPD lesion (or TT motif) in the active site of the enzyme (at $-130°$). For the CPD lesion this extra-helical conformation is of lower free energy compared to the intra-helical state by $-2\,\mathrm{kcal/mol}$ and for the TT case the extra-helical state is still less stable by $2\,\mathrm{kcal/mol}$ compared to the intra-helical state. Interestingly, the bound enzyme result in a lowering of the free energy barrier for flipping the CPD lesion or the TT bases towards the major groove but result in still a relatively high barrier for the CPD case for flipping along the minor groove. Note, that the enzyme contacts the DNA at the minor groove side which results in high sterical barriers for the bulky and rigid CPD lesion to glide along the minor groove from the intra-helical to the extra-helical state.

This is explained by steric clashes with the protein, in particular the residues TRP 391, ARG 403, and PRO 401. The CPD-damaged DNA has to change its backbone configuration to move past this obstacle. In general the internal configuration is made less stable by residue PRO 401 protruding into the WC-pair partners of the CPD base dimer (see Figure 7.7 point 1 and Figure 7.8a).

In contrast, the TT motif, although kept close to a stacked state by a few distance restraints, is much more flexible and can better adapt to the space between protein and DNA during the flipping process along the major groove. For the flipping process towards the major groove the free energy penalty is smaller even compared

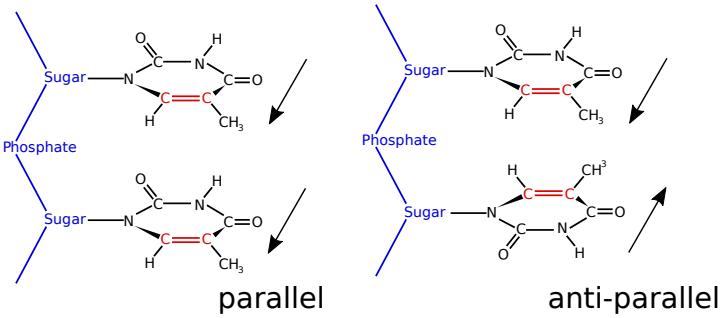


Figure 7.9: Parallel and anti-parallel orientation of adjacent thymine bases.

to the flipping with a DNA deformed towards the bound form (by approx. 1 kcal/mol–2 kcal/mol due to additional protein-DNA contacts (pushing from the minor groove and attractive interactions to move towards the active site cavity in the enzyme).

As observed in the trajectory, the **TT$^{\text{prot}}$** simulation exhibits major differences to the **CPD$^{\text{prot}}$** one. Although the two thymine bases are kept relatively close together, they are still more flexible. This allows the bases to flip around their own axis and consequently they can be oriented in opposite ways as shown in Figure 7.9. Additionally, even with restraints holding the thymine bases fairly close together, the two thymine bases are more flexible and can move more easily past the obstacle of the three previously mentioned residues of the photolyase protein in the minor groove direction.

Differences in free energy by restraining the two thymines for the simulation of undamaged DNA

It was checked whether the influence of restraining the two thymine bases together is significant or not. The US simulation was done in multiple ways. First without restraining the two thymine bases, second by restraining them with a group distance restraint and third by restraining with three distance restraint to further keep the thymine bases from switching from cis to anti conformations. As shown in Figure 7.11, the results are relatively similar. Further, the contribution of the restraint was calculated by performing a one-step Free Energy Perturbation (FEP) using the simulation without restraints. Doing this for packages of 5 windows (otherwise not enough data points where available after the fact), it is shown that the FEP by this restraint is relatively small (see Figure 7.10). The contribution of the artificial restraint between the two thymine bases was calculated, which was used for sampling reasons, with a free energy perturbation calculation. Specifically the Bennett Acceptance Ratio (BAR) method by Steffen et al. [157] was used to accomplish this with maximum accuracy. The results in Figure 7.10 show that

the perturbation by the restraints is reasonable low. It is only rarely larger than 1 kcal/mol.

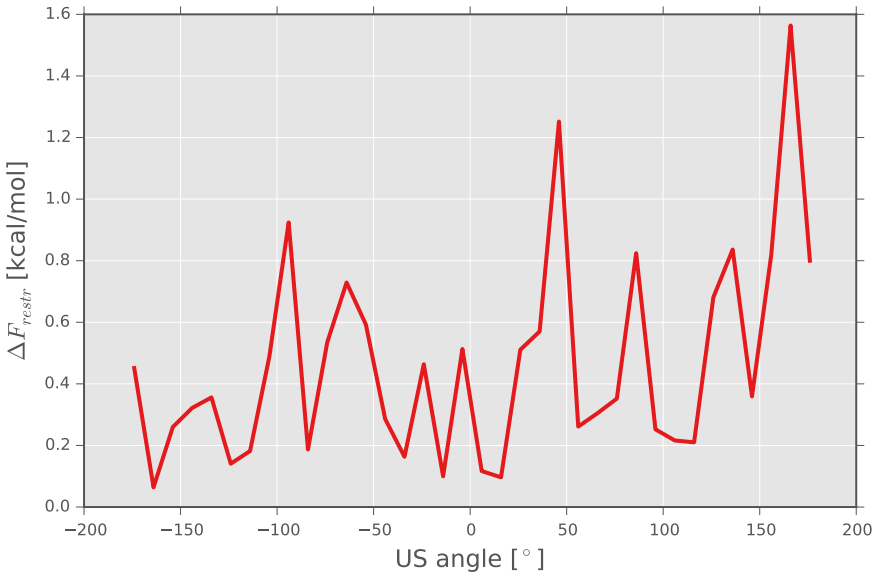


Figure 7.10: FEP contribution by the group restraint on thymine bases.

7.3.3 *Reaction Rates and Mean First Passage Times*

Simulation	τ_+ [s]	τ_- [s]	k_{eq} [1/s]	$\Delta F_{k_{eq}}$ [kcal/mol]
CPD$_{BDNA}$	3.9	5.1×10^{-8}	7.8×10^{7}	10.8
TT$_{BDNA}$	6.2×10^{1}	7.1×10^{-8}	8.7×10^{8}	12.3
CPD$_{1TEZ}$	1.1×10^{-3}	3.7×10^{-8}	3.1×10^{4}	6.2
TT$_{1TEZ}$	6.9×10^{-4}	1.1×10^{-9}	6.1×10^{5}	7.9
CPD^{prot}	3.6×10^{-6}	3.4×10^{-4}	1.1×10^{-2}	−2.7
TT^{prot}	1.4×10^{-4}	9.5×10^{-7}	1.5×10^{2}	3.0

Table 7.2: MFPTs for the flipping reaction from the intra- to the extra-helical state and vice versa.

As described in Section 2.12.6, the rate can be calculated from the intra- to the extra-helical state and vice versa using the data obtained by the US simulations. Diffusion rates are calculated as explained above. As the diffusion is a kinetic property, it converges much slower than the actual free energy. The diffusion coefficients along the reaction coordinate for the three performed types of simulations are shown in Figure 7.12, Figure 7.13, and Figure 7.14. Calculating the diffusion constants from HREMD-US simulations might be problematic as the exchanges between windows can influence the auto-correlation function. Since the largest part

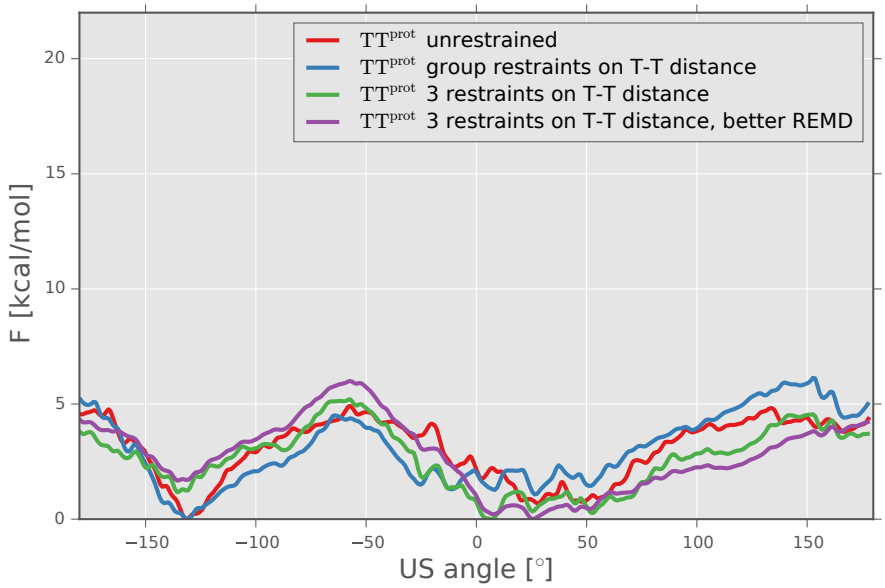


Figure 7.11: Three US simulations of $\mathbf{TT^{prot}}$. The influence of the distance restraint between the two thymine bases is negligible. This is in good accordance with the FEP calculations shown in Figure 7.10.

of the auto-correlation function declines rapidly, the longer tail is mostly interpolated and not influenced strongly by exchanges. Thus, the resulting MFPTs can only be considered estimates. These are shown in Table 7.2. The MFPTs are given as $\tau_+ = \tau_{\text{intra}\to\text{extra}}$ and $\tau_- = \tau_{\text{extra}\to\text{intra}}$. The free energy differences according to the equilibrium rates are in good agreement with the directly calculated free energy differences.

Most interestingly, the differences in the rates of the flipping reaction of the damaged and undamaged DNA (especially in the presence of the protein) are more pronounced as the free energy differences. This is mostly due to the logarithmic dependencies of the free energy of the equilibrium rate. Secondly, the shorter (around 20°) reaction pathway of the damaged DNA increases the flipping rate from the intra- to the extra-helical state dramatically. In comparison, the CPD-containing DNA can flip from the intra-helical to the extra-helical state in roughly 3 µs whereas the TT containing DNA needs about 700 µs. For similar system such as glycosylases repair proteins, proteins typically stay around 50 µs at every base site [53]. In this time, the complete repair procedure has to be finished. Any previous repair step before the flipping procedure will lower the reaction rates even further. Thus, it is very likely that only the damaged bases can flip from the intra-helical to the extra-helical state in the given time.

7.3.4 *Diffusion constant along the reaction coordinate*

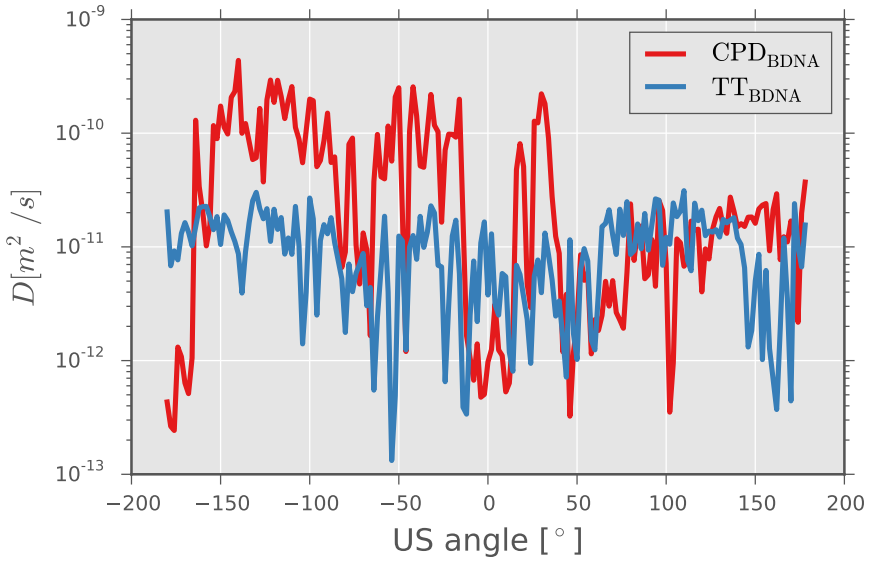


Figure 7.12: Diffusion constant along the reaction rate for the simulation of **CPD$_{\text{BDNA}}$** and **TT$_{\text{BDNA}}$**.

As discussed, the diffusion profiles along the one-dimensional reaction coordinate were calculated and are shown in Figure 7.12, Figure 7.13, and Figure 7.14. The results are fairly stable. If one takes the average diffusion constant of the diffusion along the one-dimensional reaction coordinate (see Figure 7.12, Figure 7.13, and Figure 7.14) instead, the mean first passage times almost consistently get smaller by less than one order of magnitude. The reduced roughness in the diffusion landscape increases the reaction rate, similar to the roughness of the potential as described by Zwanzig. Hence, it is shown that the used methodology is not particularly dependent on the exact profile of the diffusion coefficient.

On average the local diffusion constants are lower than expected from the simple model of Chapter 3, Section 3.2.4. As seen in Table 7.3, the maximally possible free energy barrier (shown in the last column of Table 7.3) are much lower than the anticipated 6.7 $\mathrm{kcal/mol}$ for all of the simulations. In this context, only the simulation of the DNA in complex with the protein can be realistically compared to the models given in Chapter 3. Indeed, our measured barriers are still lower than the newly predicated maximally possible free energy barriers. As the path is longer for undamaged DNA, the free energy barrier which can be surpassed is even lower 3.1 $\mathrm{kcal/mol}$ than the barrier possible for CPD-damaged DNA with 4.2 $\mathrm{kcal/mol}$. These calculations confirm the conclusions made directly from the MFPTs.

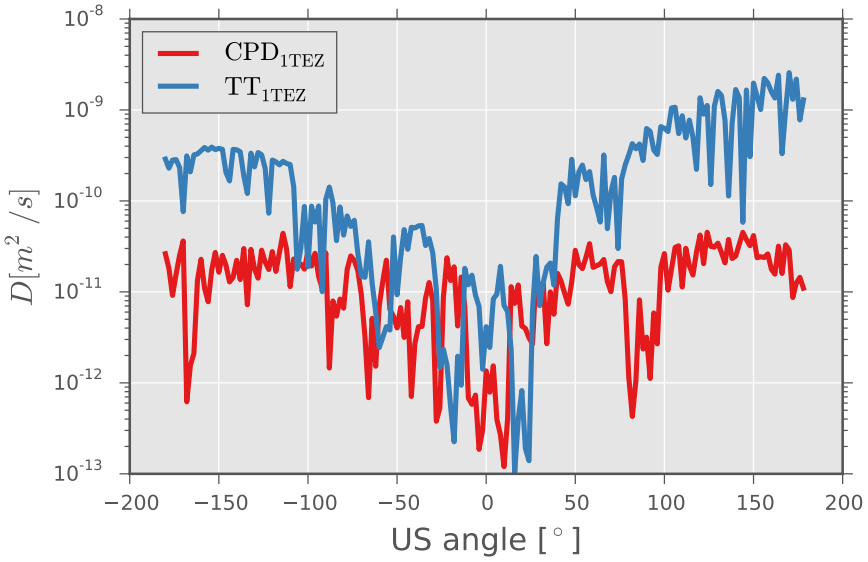


Figure 7.13: Diffusion constant along the reaction rate for the simulation of **CPD$_{1TEZ}$** and **TT$_{1TEZ}$**.

Simulation	D_{avg} [m^2/s]	ΔF_{a-max} [kcal/mol]
CPD$_{BDNA}$	5.5×10^{-11}	
TT$_{BDNA}$	1.1×10^{-11}	
CPD$_{1TEZ}$	1.5×10^{-11}	
TT$_{1TEZ}$	3.6×10^{-10}	
CPD^{prot}	8.4×10^{-12}	4.2
TT^{prot}	9.9×10^{-12}	3.1

Table 7.3: Diffusion constants averaged along the reaction coordinate.

7.3.5 *Convergence of Umbrella Sampling Simulations*

To check the US sampling simulation for convergence, the output files of the measurements of the specific reaction coordinate were split into 5 distinct or additive time intervals and calculated the PMFs for each of those time intervals. By plotting the resulting free energy curves along each other, trends and convergence can be checked for the US simulations (see Figure 7.15). Cumulative plots would in comparison by their very nature show even stronger trends of convergence.

Further on, the distributions for each of the US simulations were calculated to check whether the overlap of those distributions is sufficient. The overlap of all US distributions is approximately 40 %. For US and WHAM in particular, the method relies on sufficient overlap, meaning that the sampling of overlapping states is statistically relevant (see also [120]). Our implementation of REMD US however is a little more critical to actually achieve the benefits

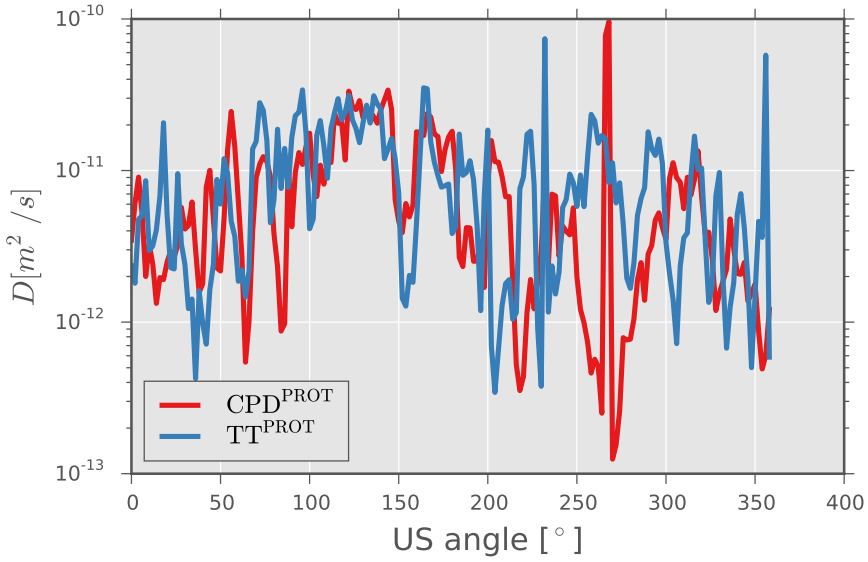


Figure 7.14: Diffusion constant along the reaction rate for the simulation of **CPD^{prot}** and **TT^{prot}**.

explained in Section 7.2.1. Optimal acceptance ratios of are generally targeted for optimal ratios of accuracy and efficiency. Some efficiency is traded in favor of accuracy with average acceptance rates of approximately 40 %.

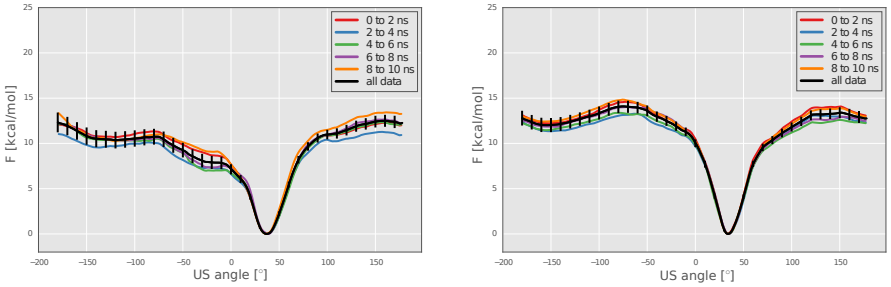
7.3.6 *Error estimation*

The implementation of WHAM[2] of Alan Grossfield includes error estimation routines. This method uses the generation of fake data sets to compute the heterogeneity of the data set [2].

As explained by Fangqiang Zhu and Hummer, The use of bootstrapping for error analysis has proven to be very unreliable due to many problems. Here the block averaging method was used accordingly to Hess. The auto-correlation time has to be calculated for each window and its time series data. Fitting of the auto-correlation function is quite hard as an approximation of the auto-correlation time with one exponential function is often insufficient. However, often fluctuations over long periods of time can occur in the time series resulting in the need of multi-exponential fits and several correlation times for different modes. Most approximations use two distinct modes. The longer mode is used if this mode is existent.

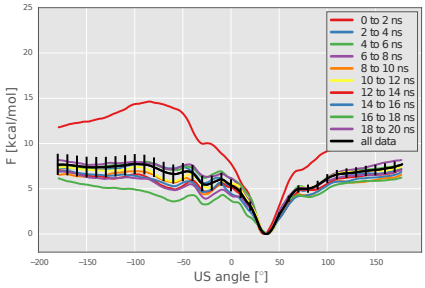
7.4 CONCLUSIONS

Comparative HREMD simulations to flip out a CPD damage or two thymine nucleobase have been performed on isolated ds-DNA molecules and ds-DNA molecules bound to a photolyase repair enzyme.

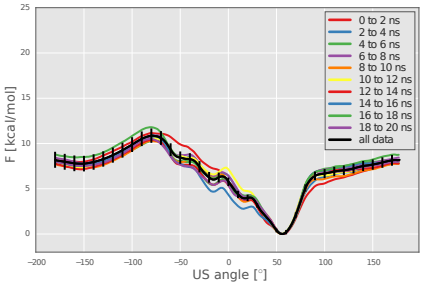


(a) **CPD$_{\text{BDNA}}$**

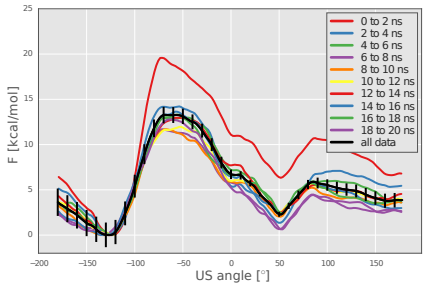
(b) **TT$_{\text{BDNA}}$**



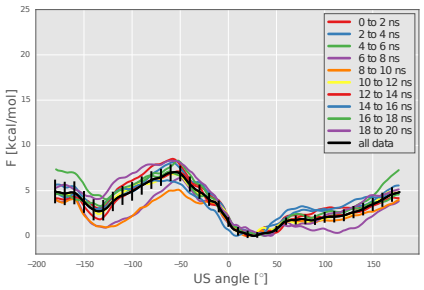
(c) **CPD$_{\text{1TEZ}}$**



(d) **TT$_{\text{1TEZ}}$**



(e) **CPD$^{\text{prot}}$**



(f) **TT$^{\text{prot}}$**

Figure 7.15: Convergence of some of the US simulations. Convergence not depicted here is similarly good. Due to slower convergence the lower four simulations have been extended to 20 ns.

The simulations indicate that in the unbound state the free energy penalty for flipping a CPD lesion from the intra-helical state to a extra helical conformation is lower compared to regular undamaged DNA. However, the free energy penalty of 10 kcal/mol is still too high to allow for frequent spontaneous flipping events. Comparing this number to the theoretical boundaries calculated in Chapter 3 Section 3.2.2 a passive recognition mechanism is highly unlikely. Under realistic assumptions and slower diffusion due to obstructions such as nucleosomes and the interrogation of undamaged DNA, the passive recognition mechanism becomes even less likely. The high free energy penalty leads to a low effective concentration of extra-helical CPD lesions relative to intra-helical states of 10^{-6}. Interestingly, the deformation of the DNA alone towards the global bound structure results already in a significant lowering of the calculated free energy penalty for flipping without involving any

direct contacts of the protein with the damaged bases. The free energy penalty toward looping out to the major groove is further lowered by the presence of the protein and contacts of protein and DNA. However, only for the CPD damage (and not the undamaged TT motif) the free energy of the extra-helical active site bound state is lower than the free energy of the intra-helical state.

Based on the simulation results the following model for damage recognition and repair can be derived. As was shown previously, CPD-damaged DNA can adopt conformations globally closer to the bound conformation compared to undamaged DNA [89] (see previous chapters Chapter 5 and Chapter 6). This explains the much higher affinity observed experimentally for the binding of CPD-containing DNA to the photolyase repair enzyme [26, 46, 50, 51, 93, 127, 146, 159, 160]. The initial binding which is partially based on conformational selection but also induced fit to deform the DNA towards a conformation that fully fits to the repair enzyme recognition surface results also in a lowering of the penalty for the flipping process. As a second step the protein facilitates the flipping of the damaged bases through the major groove. Although the barrier for this process is similar for damaged as well as undamaged DNA it is overall only favorable for flipping the CPD lesion. The photo chemical cleavage of the CPD lesion results in formation of two thymine bases that now can flip back into an intra-helical state involving only small barriers both for flipping towards minor or major grooves of the DNA. This allows an efficient product release after the repair reaction with barrier heights of less than $5\,\mathrm{kcal/mol}$. The model places the main selection step for damaged DNA at the initial phase which corresponds to recognition and the necessary DNA deformation for precise placement of the damaged DNA at the vicinity of the active site. Given the low abundances of CPD-damaged sites relative to regular TT motifs a selection at an initial step of the process is more efficient compared to a distinction at a later stage, e.g in the active site of the enzyme. After formation of an encounter complex both flipping of the CPD lesion or of undamaged TT bases is predicted to involve reduced barriers. Note, that even a flipping of an undamaged TT motif is of little consequence since it does not trigger any reaction and can easily flip back into the intra-helical state. These results are strongly supported by our estimates for the mean first passage times (see Chapter 3).

Part IV

RMSD US SIMULATIONS OF CHANGES WHILE BINDING

For a complete understanding of CPD damage recognition and repair, the most important steps of this process have to be investigated in detail. In the previous chapter the flipping process was studied. Although this process is important in the repair mechanism, it is yet still unknown if recognition completely relies on the rather small differences of flipping of damaged bases compared to undamaged bases observed in the previous chapter. As seen in Chapter 5 and Chapter 6, damaged DNA is already structurally different in comparison to native undamaged DNA. Thus, it would be beneficial for an efficient and fast repair mechanism if the damaged DNA was recognized during the encounter of the protein and the DNA damage. Using the rather novel approach of two-dimensional RMSD Umbrella Sampling with HREMD, it was shown that damaged DNA can more easily bind to the repair enzyme by assuming the conformation of the encounter complex more readily. This technique allows for a better understanding of the damage recognition process and a final proposition of the mechanism.

8

RMSD US SIMULATIONS OF CHANGES WHILE BINDING

8.1 INTRODUCTION

The recognition process was studied previously by analyzing the specificity of binding. Long (1000 ns) free MD simulations of both the damaged and undamaged structures in the absence of the protein were compared (see Chapter 6). Together with the Umbrella Sampling simulations of the base flipping step (Chapter 7), it has been shown that a passive repair mechanism is unlikely.

After binding of the DNA to the photolyase, the investigated base is flipped into the repair pocket. This was studied in detail by Umbrella Sampling simulations, first in the absence of the protein for different configurations and later in the complex with the photolyase repair enzyme.

The differences between undamaged and damaged DNA in terms of free energy profile from the intra- to the extra-helical state were not clear in all simulations of Chapter 7. The differences became clearer with the introduction of the mean first passage time as an analysis tool. For the flipping simulation in the presence of photolyase, both types of DNA, damaged and un-damaged, can flip in times in the range of microseconds. Damaged DNA can flip around 100 times faster but small errors in measurements could mean that also undamaged DNA would flip during the average stay of the protein at a DNA nucleobase. Systematic errors leading to larger free energy differences would on the other hand mean longer Mean First Passage Times (MFPTs) for both simulation types. A mechanism allowing the protein to stay longer at lesions in DNA would be needed for the damage to flip into the extra-helical position in the given time. As described in Chapter Chapter 6, global differences of the structure are present in the damaged DNA compared to undamaged DNA. These differences could lead to an efficient repair mechanism where damaged DNA binds more favorably to the repair enzyme in comparison to undamaged DNA. As lain out in Chapter 3, this process confirms the proposed third hypothesis as explained in Figure 3.3. It can be simulated how the DNA changes its conformation. This includes but is not limited to a stronger bend. The process of conformational change is depicted as sub-step 4) in Figure 3.3.

In the following it will be demonstrated this process is essential for recognition of damaged DNA. To understand this transition,

the focus in this part of the thesis will be on the RMSD Umbrella sampling of both damaged and undamaged structures from the unbound to the bound form in the absence of the protein. Hereby, the differences in binding affinity can be analyzed in more detail by not being limited to the study of the dynamics and conformations of the unbound form.

In the following will be described how to perform two-dimensional RMSD-Umbrella Sampling on the transition of a system defined by two well-defined states A and B. This method is then tested on the previously studied flipping transition of two bases (CPD versus two adjacent thymine bases). Eventually, the method is applied to the global conformational change of the DNA from the unbound to the bound conformation.

8.2 METHODS

8.2.1 *RMSD Umbrella Sampling*

In many systems it is of crucial importance to study the transition of a system from one specific conformation to another conformation. Those two conformations can often be defined by a set of particular reaction coordinates. In the cases considered in this work the transition was studied by RMSD coordinates.

Using just one RMSD coordinate with respect to state first state (A) would only define the transition to A and from A in the close vicinity of A in RMSD space. This reaction coordinate is denoted as $RMSD_A$ whereas the RMSD to the state B is denoted as $RMSD_B$. A specific distance in RMSD space from a set of coordinates does not define one particular set of coordinates. Instead it defines a subspace of m dimensions in RMSD space, more precisely an m-dimensional sphere with a radius of the RMSD distance to the first coordinate. m is given as $m = (n-1)d-1$ where n is the number of atoms and d is the dimension in space ($d = 3$ for three-dimensional space). Obviously those m dimensions are still very much correlated due to restraints on bond lengths etcetera.

Using two reaction coordinates $RMSD_A$ and $RMSD_B$, the two conformations A and B can be defined precisely without restricting the transition from A to B. Prime candidates for this methodology are transitions for which a specific coordinate such as a dihedral angle cannot be clearly defined.

An example would be the analysis of the conformational selection of DNA repair proteins. They tend to select DNA by binding to damaged DNA with a different conformation to undamaged DNA. During the binding process the protein further changes the conformation of the DNA. Thus, it is of special interest to examine the transition of undamaged and damaged DNA from the unbound

to the bound form. Multiple reaction coordinates can be found which change during this transition necessarily. Further more, for a RMSD coordinate with two references A and B, the previously explained set of reaction coordinates $(\mathrm{RMSD}_A, \mathrm{RMSD}_B)$ is also sufficient for describing the transition in terms of end points. All the atoms which change during the transition need to be included in the RMSD mask.

Umbrella Sampling, HREMD Umbrella Sampling and multi-dimensional Umbrella Sampling are time-tested methods[1]. The WHAM analysis method allows for almost arbitrary uses of different Umbrella Sampling potentials and reaction coordinates. Therefore, the best possible specific restraints can be applied to our system. As such, the RMSD provides a valid restraining solution to enhance the sampling of the specific transition in the system.

8.2.2 *RMSD-Space Sampling*

If two reference structures are present, two RMSD coordinates can be applied for the use of Umbrella Sampling. As previously noted, this allows defining the two end-states of the transition to be perfectly represented while not restricting the transition pathway excessively.

Naively, one might consider to simply do Umbrella Sampling in the whole two-dimensional RMSD space from $(0, 0)$ to $(\mathrm{RMSD}_A(B), \mathrm{RMSD}_B(A))$ where $\mathrm{RMSD}_A(B) = \mathrm{RMSD}_B(A)$. But as the RMSD is a sum of differences and therefor a mathematical norm, the triangle inequality $\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$ holds true. Therefore, the RMSD space violating the triangle inequality is not accessible and does not need to be sampled. The space to be sampled is thus the area in between $(0, \mathrm{RMSD}_B(A))$, $(\mathrm{RMSD}_A(B), 0)$, and $(\mathrm{RMSD}_A(B), \mathrm{RMSD}_B(A))$. Values even higher could be sampled but represent large detour in the transition pathway.

The associated phase space volume of a specific point in the two-dimensional RMSD space is not constant. The diagonal itself has infinitesimal small associated phase space volume as it is a one-dimensional subspace in RMSD space. Accessible phase space increases with the distance to this diagonal. On the other hand, a relatively short pathway is beneficial as it decreases the necessary distance to go from conformation A to B or vice versa.

The phase space increases with higher $\mathrm{RMSD}(A)$ and $\mathrm{RMSD}(B)$. The entropy contribution thus lowers the free energy for regions which are further apart from the before-mentioned diagonal. This should be kept in mind in the analysis of the free energy contribution and will be discussed in more detail later on.

1 Umbrella Sampling and HREMD has been successfully applied to the study of DNA damage repair (see Section 4.2) and bio-polymers in general (see Equation 2.12.5).

8.2.3 *Implementation and Computational Details*

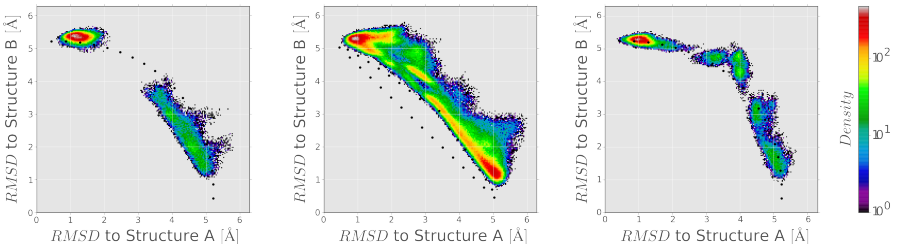


Figure 8.1: Insufficient sampling by the use of one-dimensional setup: Three of the many different tested one-dimensional methods which were used. Sampling was always insufficient. Many areas in phase space are not sampled. The x-axis describes the RMSD to structure A whereas the y-axis denotes the RMSD to structure B. The densities of sampling are shown with increasing numbers from violet to red.

If we try to sample impossible regions in the two-dimensional RMSD space, the barrier of the previously described triangle inequality will be unnecessarily over-sampled To avoid the waste of computational resources, multiple implementations of different setups were tested.

Sampling along the diagonal however covers only a region of very small phase space. I then decided to sample a circular path in the two-dimensional RMSD space as this encompassed the minimum free energy path for the systems herein tested. On a circular path the forces generated by the two RMSD restraint potentials are always perpendicular. They are completely uncorrelated. On the diagonal, one is in the regime where the two RMSD restraint potentials are completely anti-correlated. This means the reduction of one coordinate necessarily leads to the increase of the other by the same amount. In the region beyond the circular path an increase of one coordinate will go with an increase of the other coordinate. To go from conformation A to conformation B, a pathway far in the outer region is not efficient as it mostly includes large detours. But as pathways further apart from the diagonal have a higher associated phase space volume, the circular path and its surrounding region seem a likely starting point for this simulation.

It is always possible that the lowest free energy pathway is not sampled. Here, this problem would appear if the sampling region is not sufficiently bound by free energy barriers. If this is apparent in the analysis, the simulation has to be extended to accompany the required region necessary to yield the lowest free energy path.

Using a one-dimensional circular coordinate was not successful. The one-dimensional setup did not cover the regions close to the diagonal regime well enough. The theoretically infinite boundary could not be observed as it was too far away from the sampled regime.

Figure 8.1 shows examples of one-dimensional setups along the two-dimensional coordinates. The sampling cannot be controlled well enough to cover the necessary areas.

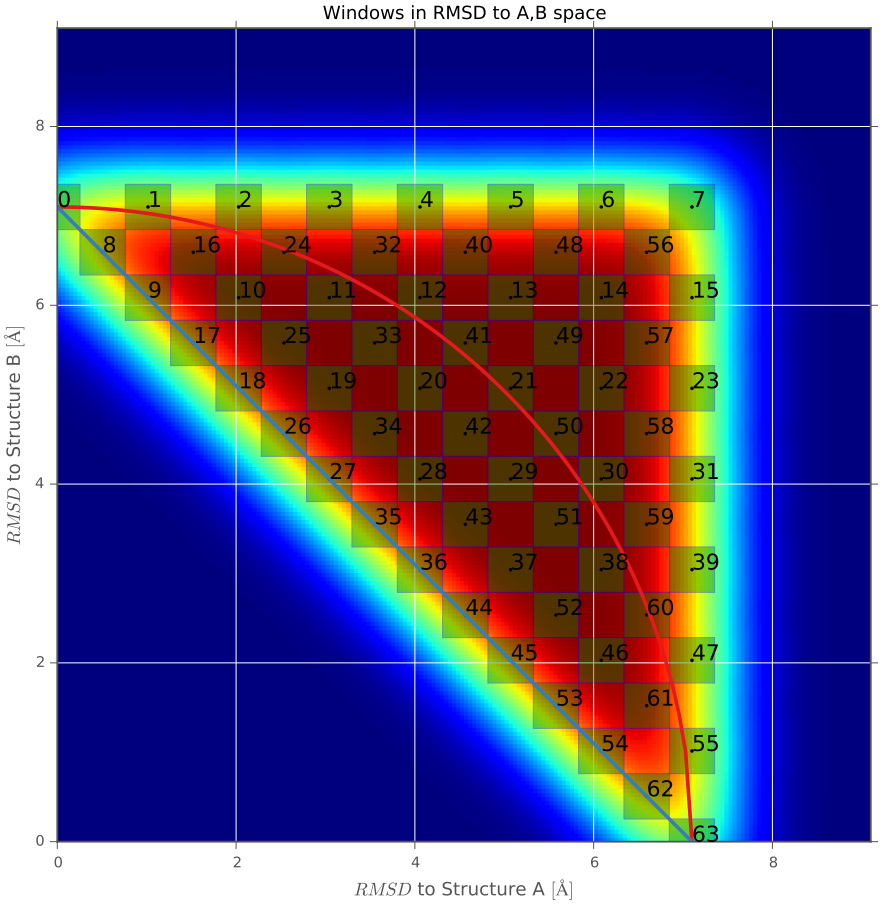


Figure 8.2: Setup and hypothetical distribution under the assumption of a flat free energy profile for two-dimensional RMSD-US. Window positions, there Gaussian widths (related to their strength) and their numbering are shown. The associated hypothetical sampling (assumption of flat free energy) is shown below the window positions.

Therefore, the methodology was changed to sample a more complete regime of the two-dimensional RMSD space. Multiple implementations were tested. In particular, by increasing the strength of sampling potentials finer details of interesting transition states can be obtained. With one-dimensional sampling a fine sampling of transition states is very difficult as it is not known before-hand where the transition states are situated. It is also possible that these transition states cover a large phase-space volume. This can be partially mitigated by two-dimensional sampling.

The complete RMSD space is divided into 8 by 8 windows, giving a total of 64 windows. These windows are located in RMSD space as shown in Figure 8.2. This specific pattern allows for the use of two-

dimensional Hamiltonian Replica Exchange MD along the Umbrella Sampling reaction coordinates. In order to use two-dimensional HREMD the number of windows in each row and column must be a multiple of 2 due to the implementation of HREMD in AMBER. The setup of Figure 8.2 is produced by laying out all windows equidistantly in a square region. The bottom-left triangle, which cannot be sampled, is then folded into the upper-right half. It is also shifted by half of the distance between windows to the bottom-left such that no regions are sampled twice. This method results in the depicted setup. Underlying the window position, the distribution related to a flat free energy profile is shown. The resulting sampling is very even (shown in Figure 8.2).

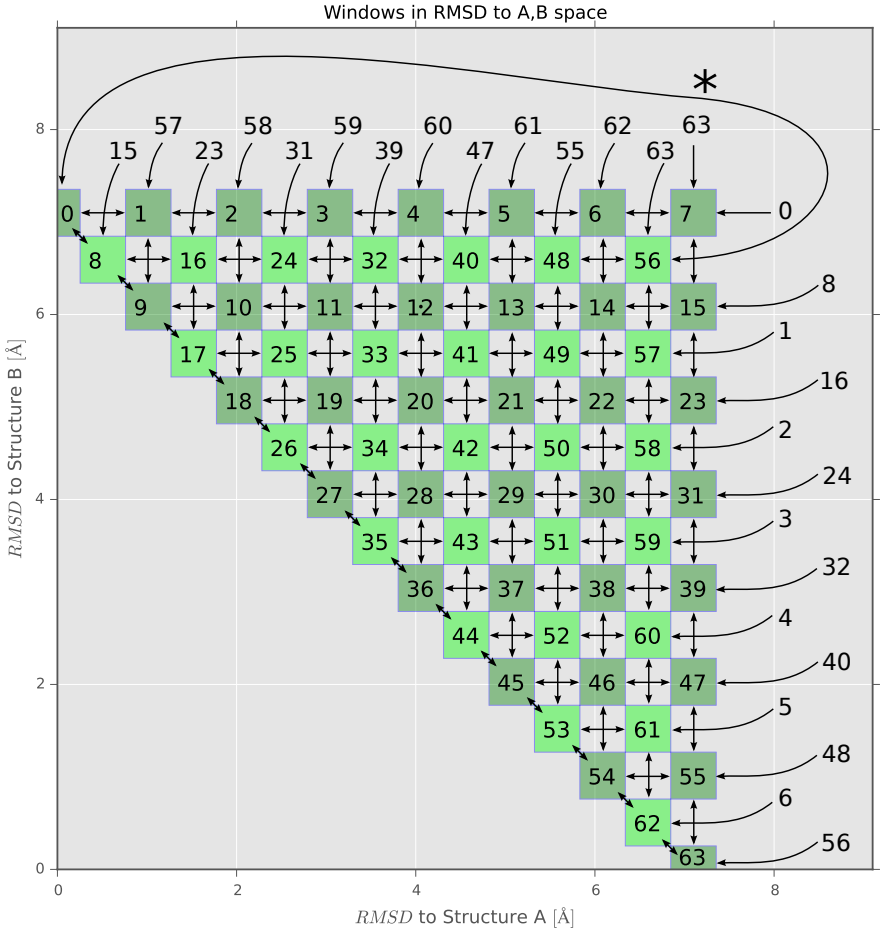


Figure 8.3: REMD exchanges for two-dimensional RMSD-US. RMSD exchanges shown with arrows. *: Some of these exchanges are hidden for the purpose of simplicity. These exchanges are explained by numbers.

The exchanges are a result of the folding procedure and are with exceptions straightforward (see Figure 8.3). The rule is that window n exchanges with the windows: $(n+1) \bmod (\lceil n/m \rceil \cdot m)$,

$(n-1) \bmod (\lceil n/m \rceil \cdot m)$, $(n+m) \bmod (m^2)$, $(n-m) \bmod (m^2)$ where the m is the number of windows in each column and row[2].

8.2.4 *Analysis Methodology*

If there are multiple visible global maxima in the resulting free energy, one may be able analyze the free energy potential and its differences directly. But as explained before (Section 8.2.2), the phase space increases exponentially with higher RMSDs in both coordinates. The exponent of this exponential increase is unknown due to unknown number of internal restraints (such as bonds) in the system. In theory this exponent could be calculated but as many inner restraints in the system are not fixed it would be a tremendously difficult task[3].

This leads to the conclusion that it is better to compare differences of these free energy curves. In particular, the resulting two-dimensional free energy surfaces are projected on a curve going from reference A to reference B. This gives a one-dimensional energy profile in return. For different systems, these can be subtracted.

As the RMSD-space very close to either structure cannot be sampled sufficiently, the difference in one-dimensional free energy has to be interpolated to the boundary cases. This will give us a comparison of probabilities of going from conformation A to B of both systems.

8.3 RESULTS AND DISCUSSION

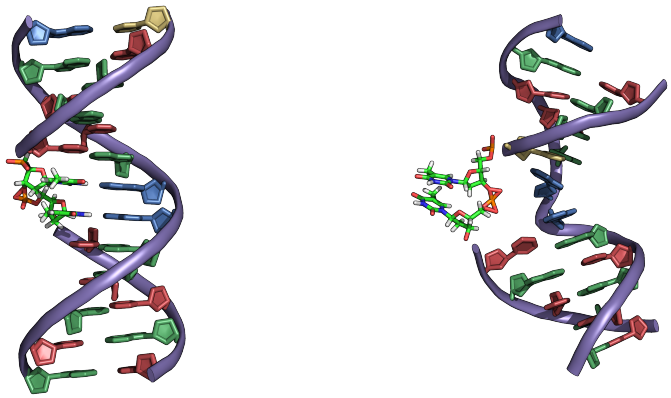
8.3.1 *Verifying the Method by Flipping of CPD*

To compare the results of this method with previous calculations, the problem of CPD flipping from the intra-helical into the extra-helical position of the duplex DNA is chosen. The results are then compared to the standard one-dimensional HREMD Umbrella-Sampling simulation results of Chapter 7. Specifically, the flipping transition without the presence of the photolyase repair enzyme is simulated. However, to set specific references A and B, two specific states which best represent the occurring biological conformations need to be defined. To compare the flipping of CPD with the flipping of two thymine bases, the same references should be selected. Otherwise, it may prove very difficult to analyses the differences further on when projecting onto a one-dimensional path and subtracting the free energy differences. As state A, perfect

2 $\lceil n/m \rceil \cdot m$ is the next multiple of m. $\lceil x \rceil$ is the ceiling, i.e. the next larger integer of x.

3 The available phase-space and it's dimension would not even be a global property but a local one as it is influenced by local properties such as available volume in the specific configuration and electrostatics.

B-DNA is selected as it represent a well-defined state. As state B, the protein bound extra-helical conformation from PDB:1TEZ is used. These two references are shown in Figure 8.4.

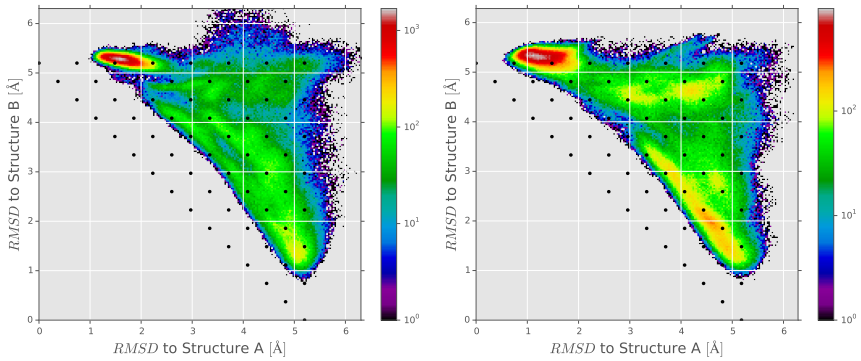


(a) Intra-helical B-DNA.

(b) Extra-helical protein bound conformation.

Figure 8.4: The two references used for the verification of the 2d RMSD-US method. Here, the undamaged DNA molecules are shown. The references of the damaged DNA are similar up to minor changes of the local structure.

Replica exchange rates are obviously not constant for all windows. However most exchange rates are well above 20 % and therefore sufficient. The exception are the exchanges which wrap around from the bottom window in one row to the top-most window. The sampling distributions of both damaged and undamaged DNA are shown in Figure 8.13. The sampling is generally quite good for both simulations. As expected, the phase space with very low phase space volume close to the diagonal does not get sampled frequently.

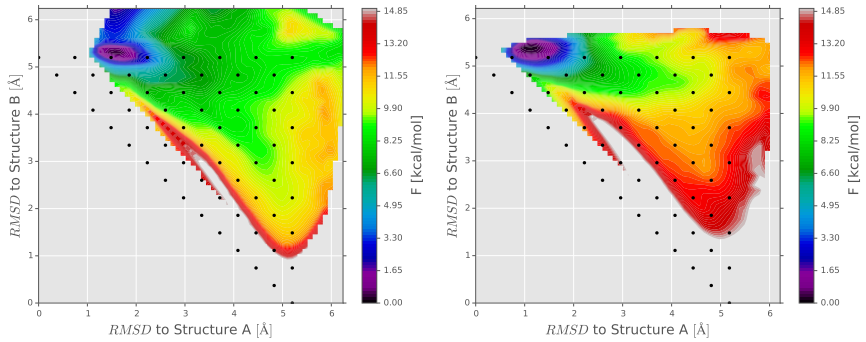


(a) Damaged DNA.

(b) Undamaged DNA.

Figure 8.5: Distribution of sampled states in simulated two-dimensional RMSD space of flipping transition.

After applying the two-dimensional WHAM procedure on these distributions and their associated window positions and strengths, the resulting free energy surface is shown in Figure 8.6.

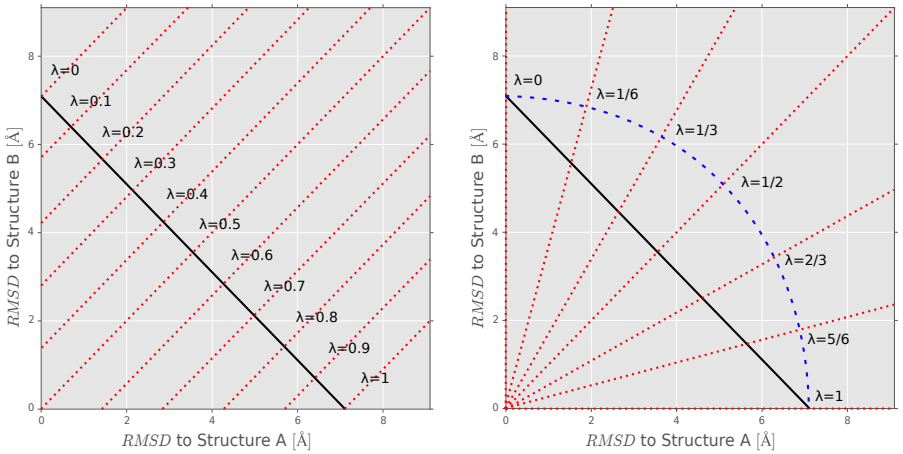


(a) Damaged DNA. (b) Undamaged DNA.

Figure 8.6: Free energy of flipping transition as a function of the RMSD to reference A and reference B. Reference A: Intra-helical B-DNA conformation. Reference B: Extra-helical conformation.

To analyze the difference of these two plots more clearly, the probabilities associated with the free energy are projected onto the one-dimensional diagonal. This was first tested by summing the probabilities up along the axis perpendicular to the diagonal (shown in Figure 8.7a). As seen in Figure 8.11, this can lead to problems as regions with very low free energy are possible that are actually further away from the reference states A or B than the orthogonal summing would imply. Therefore, summing along a radial coordinate (illustrated in Figure 8.7b) is used. The free energy is subsequently calculated from the probability. The resulting one-dimensional free energy along the diagonal is shown in Figure 8.8. The difference of these two curves (shown in Figure 8.9) has a clear trend. To get the difference for the whole transition the difference of free energy for the value of $\lambda = 1$ has to be calculated. As the difference for this value cannot be calculated, it must be interpolated. As there is no model for the difference in free energies, a linear fit is the first guess. Indeed the difference can be fitted over a long range of λ.

Using this linear fit on the whole range of λ shows that the free energy barrier is at approximately $4.5(15)\,\mathrm{kcal/mol}$ smaller for damaged DNA than for its undamaged counterpart. The error was calculated using standard uncertainty propagation of the fit uncertainties. Both the absolute values as well as the differences are in good agreement with previous results from Chapter 7. It can be concluded that the current method can measure the difference in free energy for reasonably simple pathways.



(a) Summing of the probabilities orthogonal to the diagonal.

(b) Radial summing of the probabilities onto the one-dimensional coordinate. A one-dimensional coordinate which gives such equidistant radial lines would be the circular line (blue, dashed).

Figure 8.7: Multiple choices are possible to sum the probabilities from to-dimensional space onto a one-dimensional coordinate. The red, dotted lines indicate examples in which direction the probabilities are summed.

8.3.2 *Transition of Damaged and Undamaged from B-DNA to the Protein Bound Conformation*

Now, the current method is used to simulate the transition of damaged and undamaged DNA from B-DNA (Figure 8.10a to) to the protein bound conformation (Figure 8.10b). Therefore, the references A and B are defined as the B-DNA structure and the protein bound conformation, respectively.

Comparing Figure 8.11 to Figure 8.12, it can be seen that the minimum of the damaged structure is closer to the protein bound form than for the undamaged structure.

For both simulations, the free energy profile does not present any sampling problems and is generally quite smooth. The sampling is sufficient and even regions of high RMSD are well sampled (see Figure 8.13).

As before, the two-dimensional free energy is projected onto the one-dimensional coordinate connecting states A and B. The result is shown in Figure 8.14. For the CPD-containing DNA, the barrier for the change in conformation is clearly lower compared to undamaged duplex DNA. The difference is plotted in Figure 8.15. Structural change is favored by at least $2.3(6)\,\mathrm{kcal/mol}$ according to the fitted curve. The original curve shows that the fit does not work for structures close to the final protein bound conformation. Thus, the

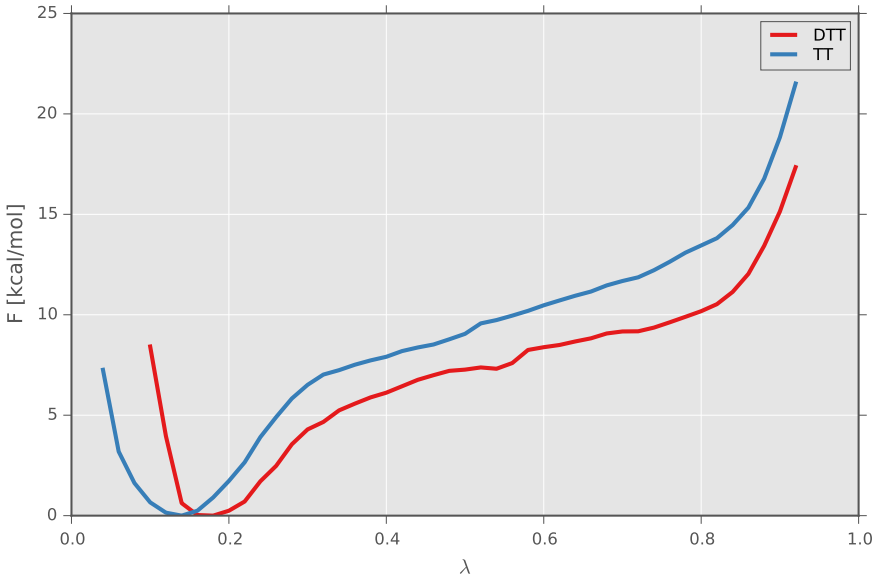


Figure 8.8: Free energy of attachment transition as a function of the RMSD to reference A and reference B. Reference A: Intra-helical B-DNA conformation. Reference B: Extra-helical protein bound conformation.

previously noted difference of 2.3(6) kcal/mol can be taken as a lower limit.

8.3.3 *Analysis of Convergence*

To analyze the convergence of the previously mentioned results, the simulation is split into multiple parts, in particular 5 parts. Then, the projections onto the one-dimensional coordinate are compared. The results are shown in Figure 8.16. By splitting the simulation into 5 pieces, no trend can be observed. Thus, the conclusion can be made that the simulation is in the converged regime.

8.4 CONCLUSIONS

Using the relatively novel application of a two-dimensional RMSD coordinate on HREMD Umbrella Sampling, it was shown that the damaged DNA structure more easily adapts transient conformations resembling the protein bound complex conformation. More precisely, this preference was measured to be at least 2.3 kcal/mol from the DNA's native structure. Assuming that the repair enzyme spends a limited time interrogating the possible damaged DNA, it is of key importance that damaged DNA undergoes the sampled transition into the protein bound conformation. It is not necessary to flip the damage into the external configuration beforehand. Even a relatively small difference in the binding affinity of damaged DNA compared

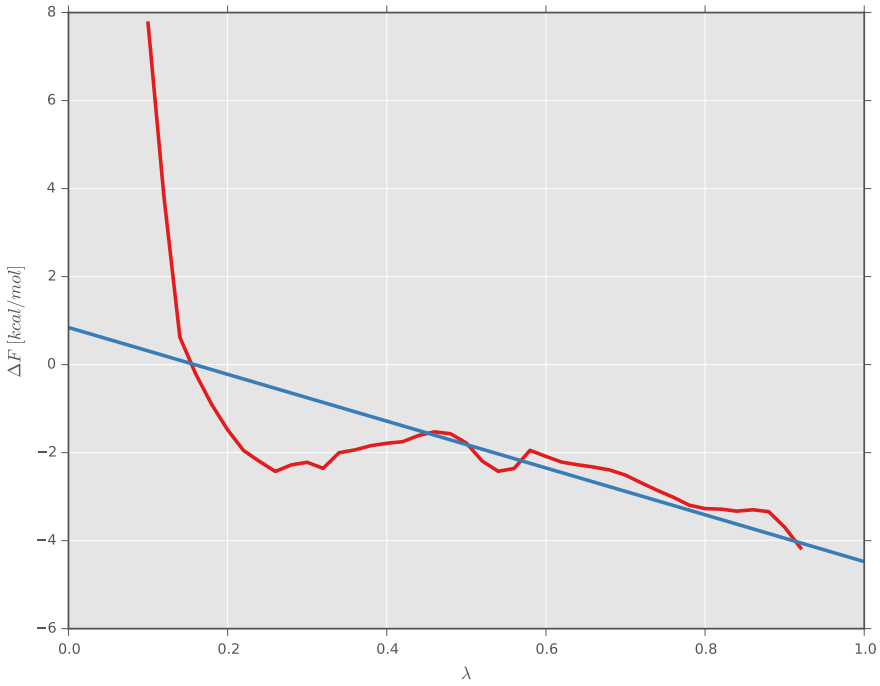
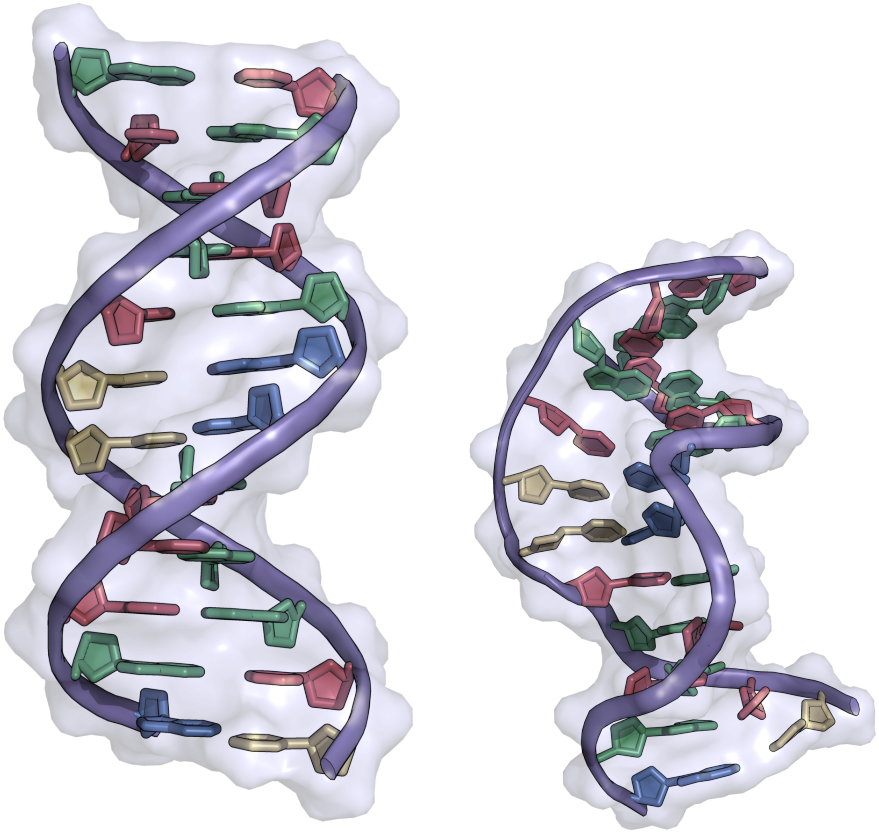


Figure 8.9: Difference of one-dimensional projection of the free energy of the flipping transition of undamaged and damaged DNA.

to undamaged DNA can benefit the repair mechanism. These results favor strongly the mechanism explained in Chapter 3, Section 3.2.6. A more complete and comprehensive conclusion and outlook will be given in the following chapter.



(a) B-DNA conformation of the DNA. Some simulations were started from this structure. The undamaged structure is labeled as $\mathbf{TT_{BDNA}}$ and the damaged structure as $\mathbf{CPD_{BDNA}}$.

(b) Deformed DNA structure by restraining all backbone atoms (except the damaged nucleotides) to the CPD-damaged DNA crystal structure in complex with the E. coli photolyase (pdb-entry: PDB:1TEZ [117]). The undamaged structure is labeled as $\mathbf{TT_{1TEZ}}$ and the damaged structure as $\mathbf{CPD_{1TEZ}}$.

Figure 8.10: The two reference structures used for RMSD Umbrella Sampling.

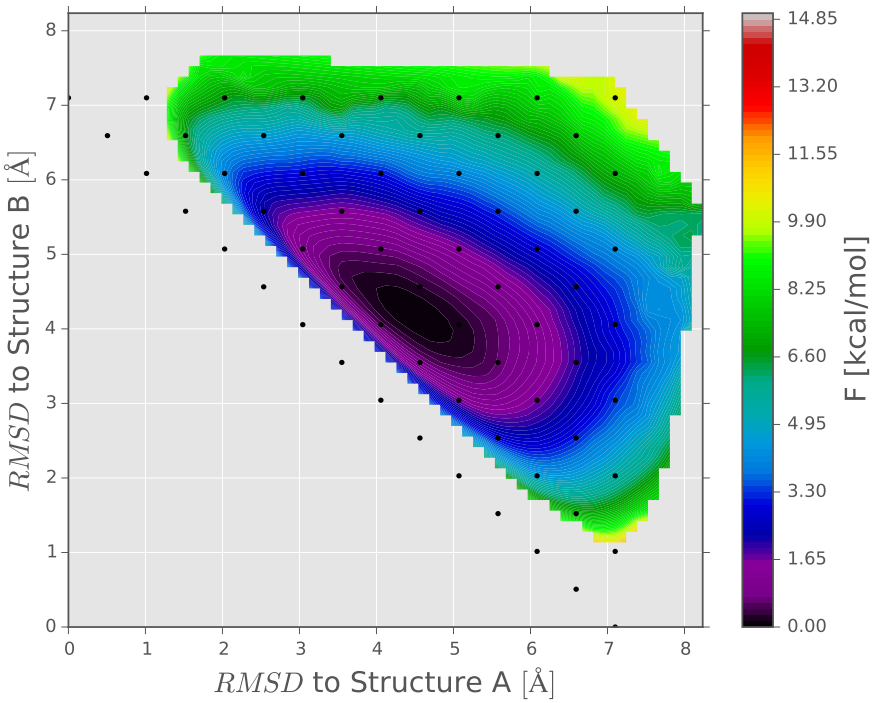


Figure 8.11: Free energy of attachment transition of damaged DNA as a function of RMSD to reference A and reference B. Reference A: B-DNA conformation. Reference B: Protein bound conformation.

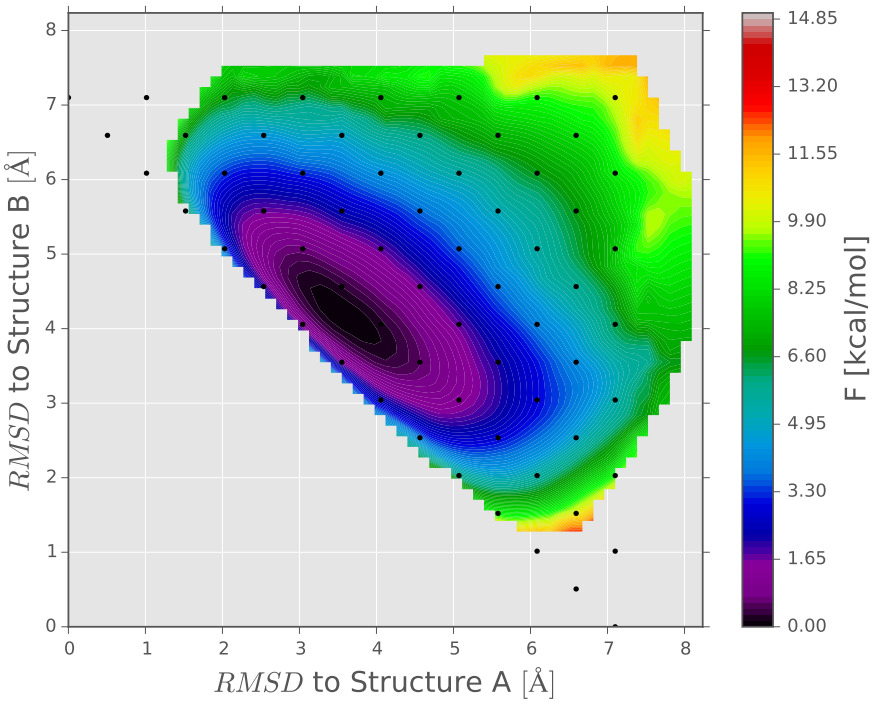
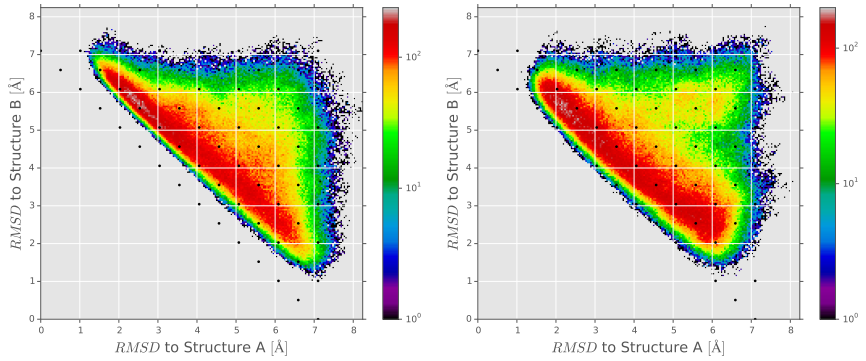


Figure 8.12: Free energy of attachment transition of undamaged DNA as a function of the RMSD to reference A and reference B. Reference A: B-DNA conformation. Reference B: Protein bound conformation.



(a) Damaged DNA. (b) Undamaged DNA.

Figure 8.13: Distribution of the sampled states in two-dimensional RMSD space in the simulation of the attachment transition.

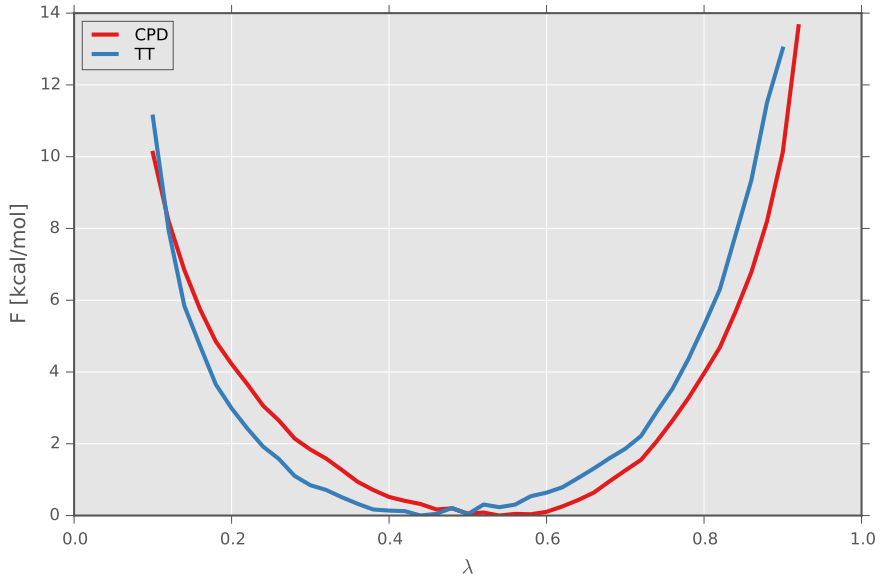


Figure 8.14: Free energy of attachment transition as a function of the RMSD to reference A and reference B. Reference A: B-DNA conformation. Reference B: Protein bound conformation.

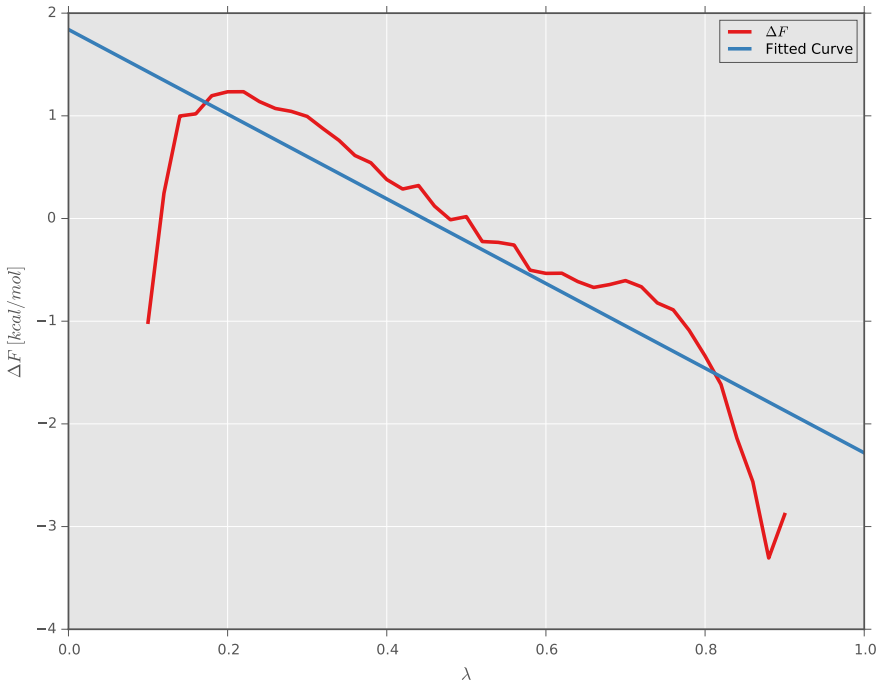
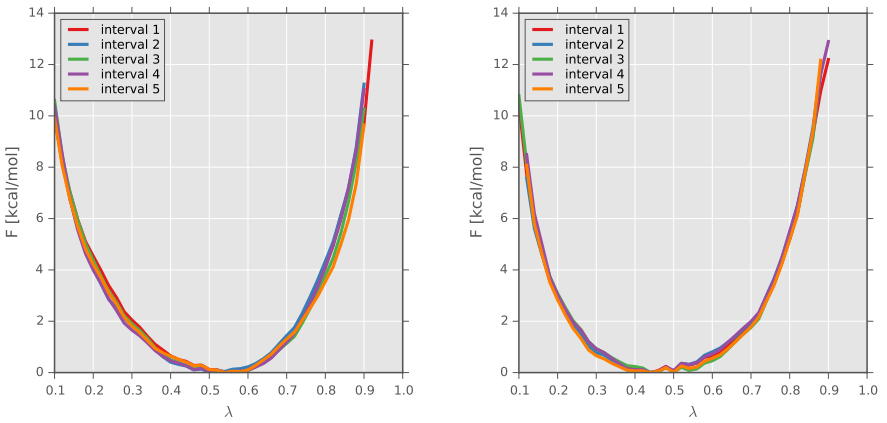


Figure 8.15: Differences of free energy of attachment transition as a function of the RMSD to reference A and reference B. Reference A: B-DNA conformation. Reference B: Protein bound conformation.



(a) Simulation of damaged DNA. (b) Simulation of undamaged DNA.

Figure 8.16: Convergence of 2d RMSD US simulations of the attachment transition.

Part V

CONCLUSION

In the last part of the thesis the results will be summarized and a conclusion given. A short outlook of what could be done in the future is given.

9 CONCLUSION

9.1 DISCUSSION

Great insight about the mechanism of Cyclobutane Pyrimidine Dimer (CPD) damage repair was obtained during the research of this thesis. The understanding of the repair and recognition process of CPD lesions in Deoxyribonucleic Acid (DNA) has been greatly increased. Some simpler scenarios were ruled-out as they strongly disagree with the results of performed simulations.

All evidence of experimental results (see Section 4.1), previously obtained Molecular Dynamics (MD) studies (see **??**) and the MD studies conducted for this work (Chapter 5, Chapter 6, and Chapter 7) propose that a passive recognition mechanism as explained in Chapter 3 is very unlikely. In detail, none of the simulations revealed any spontaneous flip from the intra- to the extra-helical conformation despite the long simulated time of $1000\,ns$[1] (Chapter 6). Neither gave any other simulation evidence of such processes. Further, the direct calculation of the free energy barriers of flipping in bulk (Chapter 7, Section 7.3.1) predicted barriers which are approximately $10\,kcal/mol$. For the simple model of passive recognition presented in Chapter 3 the free energy differences would need to lower than $8.9\,kcal/mol$. As this model strongly overestimated the diffusion in bulk and underestimates the influence of the DNA on the search of the protein, even lower boundaries are expected for the passive recognition mechanism. More precise assumptions were included in the revised models in Section 3.2.3.

The second described model proposes a recognition mechanism during the flipping stage of repair. In that case, the repair enzyme attaches subsequently to every base of the DNA one after another. Many bases will be checked multiple times as the sliding of the repair enzyme acts happens in a manner of a random walk (see Section 3.2.3 and [53]). Comparisons with other repair enzymes such as Glycosylase suggest that the protein visits one base for an average time of $50\,\mu s$ [53] (see Section 3.2.5 for more details). In theory, all bases could be checked whether they flip into the repair site or not, provided that the binding process of the DNA to the protein is fast enough. If the DNA is in the conformation close the repair protein but still intra-helical, the US simulations, which included the

1 Such simulation times have only been possible in recent years. By using Graphics Processor Unit (GPU)-acceleration, the speed of the simulation was increased from a $5\,ns/day$ to approximately $30\,ns/day$ resulting in a complete run-time of roughly a month.

protein, can accurately measure and predict the repair mechanism. This mechanism relies on the close binding of the protein to every base of the DNA. However, it may be that the conformational changes required for close binding do not happen by themselves in reasonable times, that is in the order of few microseconds, if no damage is present.

The four types of Umbrella Sampling (US) simulations which were performed during this thesis can shed light on the detailed steps involved in the repair. As expected from the results described in the Chapter 5 and Chapter 6, the global conformation of the DNA has a strong influence on the flipping mechanics. By constraining the DNA to the bend and twist of the protein bound structure[2] the free energy barrier was reduced by $2\,\mathrm{kcal/mol}$. Only by comparing that simulation to another restrained simulation (B-form DNA (B-DNA) restrained) it could be seen that the effect is even larger, specifically in the range of $5\,\mathrm{kcal/mol}$. Apparently, the restraining is helpful but constricts the flipping of the respective bases in the DNA too strongly.

A more realistic scenario was simulated by including the protein. Here, also local interactions between the protein and DNA are included. In detail, it was observed that a proline helps break up the Watson-Crick (WC) bonding of the dimer with its partners. The specific mechanism involves the side chain protruding slightly into the space between the adenine bases opposite to the possibly damaged thymine bases. By pushing them apart slightly, the breaking of the WC pair bonds is promoted. Further, the residues of a triptophane and an arginine prohibit the base flipping through the minor groove. Large differences in the structural properties and the dynamics of the bases between the thymine dimer CPD and the comparison model of two adjacent thymine bases were found. In the studies describes in Chapter 6 it was observed that the intra-helical native state of the damaged base is slightly opened in comparison to undamaged DNA. This opening is further increased in the presence and close contact of the DNA and the protein. Specifically, the angle as measured by the pseudo-dihedral angle defined in Section 7.2.1 is increased from 30° (native DNA) to 53° for CPD-containing DNA. Together with a lower free energy of the extra-helical state for damaged DNA and a slightly lower free energy barrier, the mean first passage time of CPD from the intra- to the extra-helical state in the presence of the repair protein is significantly lower than of undamaged DNA thymine bases.

A third model describing the recognition mechanism was presented in Section 3.2.6 which can even better explain the obtained results. As the flipping from the intra- to the extra-helical state is not the first process in active damage recognition and repair, it must

2 The DNA was restrained to the DNA structure present in the interrogation complex with the protein.

be checked, if any previous steps can distinguish between damaged bases and undamaged bases. Therefore, the binding process of DNA to the repair protein must be studied in detail. Measurements of sliding length have used obstructions to determine the average length of repair protein sliding along DNA [53]. It cannot be measured exactly how close the sliding is, whether the protein changes its conformation during sliding, if and how it attaches to damaged or undamaged base sites. As the protein binds strongly to DNA and does not seem to be sequence specific [117], it must be studied how the DNA adopts the binding conformations and if any free energy barriers have to be overcome. This was studied with the method introduced in Chapter 8.

As observed in Chapter 6, CPD-containing DNA adopts transient conformation which are closer to the protein-bound state. The conformational changes, which have to be overcome, are therefore of smaller magnitude for CPD-containing DNA. This explains the much higher affinity observed experimentally for binding of CPD-containing DNA to the photolyase repair enzyme [26, 46, 50, 51, 93, 127, 146, 159, 160].

Indeed, using the rather novel application of Umbrella Sampling on a two-dimensional Root-mean-square deviation (RMSD) coordinate, the transition from the native to the protein-bound state was measured to be significantly lower, specifically 2.3 $\mathrm{kcal/mol}$. This difference leads to very different characteristics of damaged and undamaged DNA in the vicinity of the repair enzyme. As stated in the third model of Chapter 3, the recognition of the repair enzyme relies on the configurational differences of damages in DNA sequences.

To summarize, the following model for damage recognition is most likely in accordance with the obtained results.

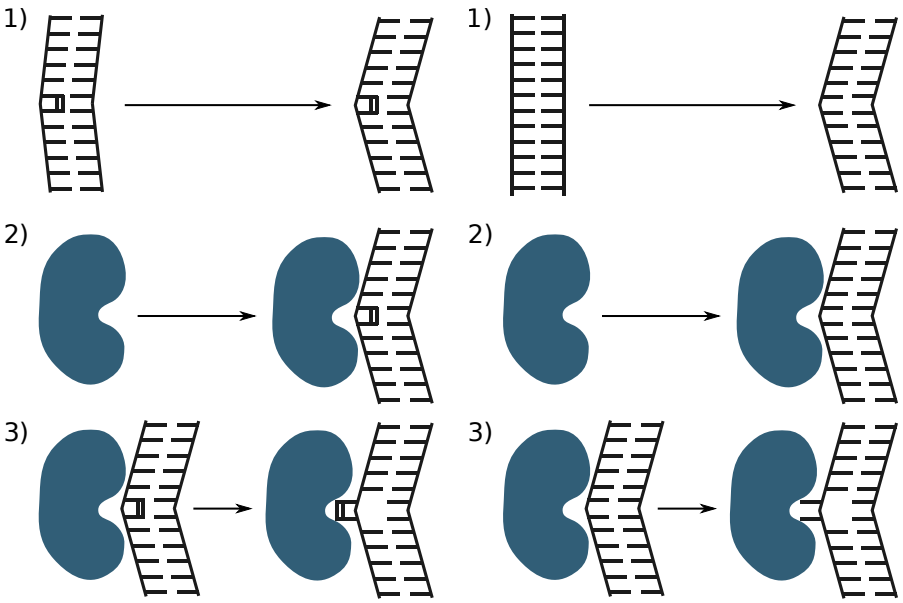
> *"Once you eliminate the impossible, whatever remains, no matter how improbable, must be the truth."*
> *Sherlock Holmes - Arthur Conan Doyle*

The initial binding is partially based on conformational selection but also on induced fit to deform the DNA towards a conformation that fully fits to the repair enzyme recognition surface. This results in a lowering of the free energy penalty for the flipping process of investigated bases as a second step the protein facilitates the flipping of the damaged bases through the major groove. Although the barrier for this process is comparable for damaged as well as undamaged DNA, it is overall only favorable for the flipping of the CPD lesion.

The photo-chemical cleavage of the CPD lesion results in formation of two thymine bases that now can flip back into an intra-helical state involving only small energy barriers both for flipping towards minor or major grooves of the DNA. This allows an efficient product release

after the repair reaction with barrier heights of less than 5 kcal/mol. The model places the main selection step for damaged DNA at the initial recognition phase which requires DNA deformation for precise placement of the damaged DNA at the vicinity of the active site. Given the low abundances of CPD damaged sites relative to regular Thymine Thymine adjacent pair (TT) motifs a selection at an initial step of the process is more efficient than a distinction at a later stage, e.g in the active site of the enzyme. After formation of an encounter complex, both flipping of the CPD lesion or of undamaged TT bases is predicted to involve reduced barriers. However, at this stage, the CPD motif flips in a time of a few microseconds into the active site of the protein whereas the TT motif requires a longer waiting time of a few hundred microseconds. Note, that even a flipping of an undamaged TT motif is or little consequence since it does not trigger any reaction and can easily flip back into the intra-helical state.

9.2 OUTLOOK



(a) Repair mechanism for the probing at the damaged site.

(b) Repair mechanism for the probing at the undamaged site.

Figure 9.1: The steps of the repair mechanism are divided into submechanisms which can be simulated in MD. The order of the steps is not necessarily in accordance with the repair mechanism.

As demonstrated in the previous chapters, the models lain out in the beginning of this work in Chapter 3 could not all be completely disproved. For a better falsification of model 2, longer and more sophisticated MD simulations are needed. Nevertheless, it could be shown that the mechanisms of model 3 will take effect.

In the following, the particular simulation protocol required for the falsification of hypothesis 2 and solidification of model 3 will be given. As mentioned in Chapter 3, the overall free energy difference between the bulk state and the protein bound extra-helical state has to be measured. Here, only parts of these mechanisms could be simulated. The steps of the mechanism were divided into sub-mechanisms, which could be simulated in MD. The order of the steps is not necessarily in accordance with the repair mechanism.

For binding of DNA at the probed site, DNA undergoes a transition to the conformation which resembles the DNA bound to the repair enzyme (see Figure 9.1, step 1). This step will happen during the binding mechanism itself but is separated from binding for ease of simulation. The second sub-mechanism is the binding of the DNA to the protein while being already in the protein bound conformation (see Figure 9.1, step 2). The third sub-mechanism is the flipping of the probed bases (see Figure 9.1, step 3). Step 1 and 3 have been simulated. Step 2 has not been simulated as it is believed that the interactions of the DNA and the repair protein are not sequence-specific. Further, as the protein binds with high affinity to the DNA, this steps was thought to be of lesser importance with regards to a free energy barrier. A Umbrella Sampling simulation using the center of mass distance as a reaction coordinate could answer the question of free energy barrier and difference of step 2. However, the effort for the simulation is immense as very large bounding boxes and consequently large number of water molecules would need to be simulated[3]. The number of atoms could easily grow by a factor of 2, increasing the simulation time drastically. Here, a cuboid periodic boundary box was assumed as it is often the most efficient method for distance Umbrella Sampling simulation. The position and orientation of the protein have to be restrained in order to use a cuboid periodic box.

With the ever increasing computing power - at least for now - even more ambitious projects in MD simulation can be started. With current available resources and computation performance, a continuous MD simulation of the whole repair process is out of question. One might argue that even with enormous amounts of computing power a continuous MD simulation of a process in the range of hundreds of micro-seconds is not an efficient use of computing resources as most of the computation time is spent in the lowest energetic states. Like Umbrella Sampling, other methods have been invented which are not as wasteful as continuous MD but recover some lost information of US and similar methods such as Hamiltonian Replica Exchange Molecular Dynamics (HREMD). As such, US cannot give us continuous trajectories and other kinetic information such as reaction rates.

3 In the current system, 17947 water molecules needed to be simulated.

In the future, the whole sliding process (and later even hopping) of the repair protein along the DNA will be possible to be simulated. These processes are in the range of milliseconds[4]. The technical challenges are enormous even with the use of techniques such as Umbrella Sampling as the size of the system and the available phase space is very large. At the moment, such simulations are not feasible. Nevertheless, I am certain that advances in computational power and methodology will allow the study of the recognition and repair of DNA damaged as a whole in the future. Up to this point only the simple repair processes involving bacterial photolyase, specifically those of Escherichia coli (E. coli), were investigated. The even more complicated system of human CPD damage recognition is especially worthy to be studied and will present further interesting challenges.

"Things are only impossible until they're not!"
Jean-Luc Picard - Star Trek: The Next Generation

4 Hopping happens approximately every 50 µs pushing the whole simulation in the range of hundreds of microseconds to few milliseconds for better statistics.

Part VI

APPENDIX

In the appendix the set-up method will be explained once such that the main chapters do not need to explain it multiple time. Subsequent are the bibliography and the acknowledgments.

A

METHODS OF SETUP AND SIMULATION

A.1 GENERAL SETUP

Here, the general setup of the simulation will be explained. All steps which are similar are explained here instead of the specific chapters. The specifics of each different type of simulation will still be covered in each of the chapters.

All MD simulations were performed in explicit water (TIP3P) [79] with a truncated octahedral box and a minimum distance of 10 Å between solute and box boundary. Potassium ions and chloride ions were included to neutralize the system and to adjust to physiological salt concentration to approximately 100 nM. For more accurate ionic parameters the Joung/Cheatham ion parameters for TIP3P water [80, 81] were used. The simulations were carried out with the Particle Mesh Ewald Molecular Dynamics (PMEMD) module of the Amber12 package using the ff99bsco and chi.OL3 force fields for nucleic acids [135, 184] and the ff99SB for proteins. The parameters by Spector et al. were used to describe the CPD damage [153].

The simulation systems were first energy minimized (1500 steps) and heated (each step 100 ps) to 300 K in three steps of 100 K followed by gradual removal of the positional restraints from 25 $kcal/(mol\,Å^2)$ to 0.5 $kcal/(mol\,Å^2)$ (in 5 steps) and a 1 ns unrestrained equilibration at 300 K. In order to avoid fraying of the terminal base pairs and to mimic the embedding of the oligonucleotides, i.e. short DNA molecules, in the context of longer double strand DNA, distance restraints to keep the first and last base pairs in a hydrogen-bonded geometry were employed.

All MD simulations were performed using the Amber12 suite of programs [27]. At later stages of the thesis, Amber14 was used in order to improve performance of calculations. However, the same force fields were used as before.

A.2 DETAILS OF THE 1TTD-SYSTEM SETUP AND SIMULATION

For the comparative simulation of the backbone angles, the crystal structure of PDB:1TTD was used as a starting point (see Figure A.1 [113]). The crystal structure contains CPD-damaged double stranded DNA (ds-DNA) with the sequence of d(5'- GCACGAA|CPD|AA-G/5'-CGTGCTTAATTC) (cyclobutane pyrimidine dimer abbreviated as |CPD|). As a comparison, a B-DNA structure with the same undamaged sequence of d(5'-GCACGAATTAAG/5'-CGTGCTTAA-

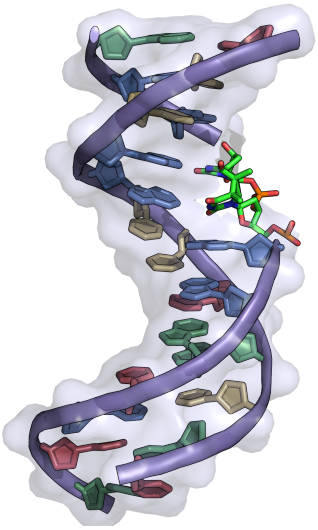


Figure A.1: Crystal structure of PDB:1TTD [113] used for the comparative analysis of the backbone angles.

TTC) was created with Nucleic Acid Builder (NAB) tool of the Amber12 tools suite [27]. It was then simulated for the same amount of simulation time as the damaged DNA. The sequence of the undamaged DNA is illustrated in Figure A.2) in which the gray box represents the position of the CPD damage. The analysis was done using the *ptraj* tool of the Amber12 tools suite and subsequent statistical analysis using python, *scipy* [78], *numpy* [168], and the *pandas* [114] libraries . The graphs were plotted with the use of the *matplotlib* library [74].

A.3 DETAILS OF THE 1TEZ-SYSTEM SETUP AND SIMULATION

A majority of the simulations were done using the system derived from the crystal structure of PDB:1TEZ [117]. The simulated DNA sequence is identical to the DNA sequence of the CPD-damaged DNA in the crystal structure of the complex with E. coli DNA-photolyase (PDB:1TEZ [117] - Figure A.3). The sequence of the duplex DNA was d(5′-TCGGCTTCGCGC/5′-GCGCGAAGCCGA) for the undamaged regular DNA. By replacing the central two thymine base with a cyclobutane pyrimidine dimer (termed |CPD| here), the sequence of d(5′- TCGGC|CPD|CGCGC/5′-GCGCGAAGCCGA) was obtained. Figure A.4 represent these two sequences wherein the gray box represents the position of the CPD damage.

B-DNA starting structures of the isolated undamaged and damaged DNA duplexes were generated with the NAB tool of the Amber12 tools suite [27]. These structures are named as $\mathbf{TT_{BDNA}}$ and $\mathbf{CPD_{BDNA}}$, respectively, and contain the central bases in an intra-helical conformation. The undamaged B-DNA structure is shown

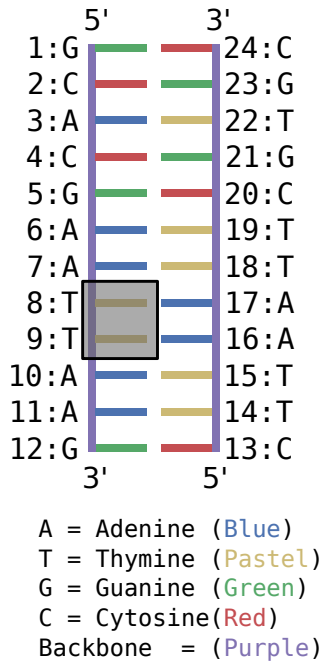
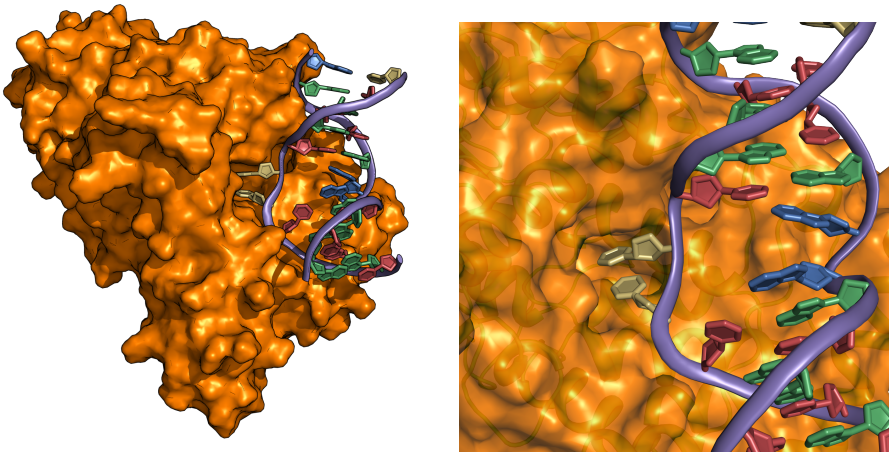


Figure A.2: Sequence of the modified PDB:1TTD [113] crystal structure. Grey box marks the position at which the crystal structure contains the CPDlesion. For comparison, a regular DNA was simulated by replacing the CPDlesion with two thymine bases.

in Figure A.5a. In order to control the influence of the starting structure on the simulation results, a second set of simulations was started from DNA conformations close to the enzyme bound form but intra-helical central CPD damage. The structures were generated by a short simulation (1 ns) starting from the B-DNA conformations including tight positional restraints referenced to the heavy atoms of the DNA structure in the complex with the photolyase (in PDB:1TEZ), but excluding the extra-helical bases and directly flanking nucleotides. This procedure gives a new starting structure with an open minor groove and similar bending as seen in the enzyme bound DNA-structure but with the central bases still in an intra-helical conformation. The start structures are termed $\mathbf{TT/CPD_{1TEZ}}$. The resulting deformed structure is shown in Figure A.5b.

Finally, MD simulations were also initiated from the DNA in complex with the DNA-photolyase repair enzyme and the central bases in an extra-helical conformation bound to the enzyme active site corresponding to the PDB:1TEZ structure [117]. The dangling ends in the experimental structure were replaced with the same terminal nucleotides as used for the simulations in the absence of the enzyme. The resulting structures are termed $\mathbf{TT^{prot}}$.

The equilibrated structures served as starting structures for production simulations of 1000 ns for the DNA in the absence of the repair enzyme and 600 ns for the complex. The complete system of the complex consists of 474 protein residues, one Flavin Adenine



(a) Full view. (b) Close up view.

Figure A.3: Representations of the slightly modified structure of PDB:1TEZ [117]. The DNA is in a complex with the photolyase protein (orange surface representation).

Dinucleotide (FAD) residue and 24 DNA bases. All systems were neutralized with 22 potassium ions. The system of the complex consists of 8321, that is: 84 atoms in FAD, 757 atoms in the DNA, and 7480 atoms in the protein. The system of the complex has in total 61942 atoms including 17947 water molecules. The smaller system of the isolated DNA consists of 16979 atoms, thereof 757 atoms for the DNA and 5102 water molecules.

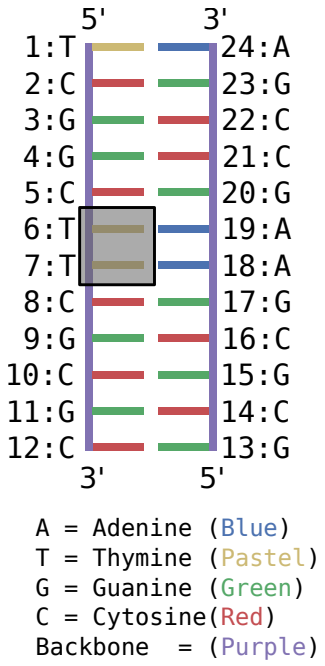
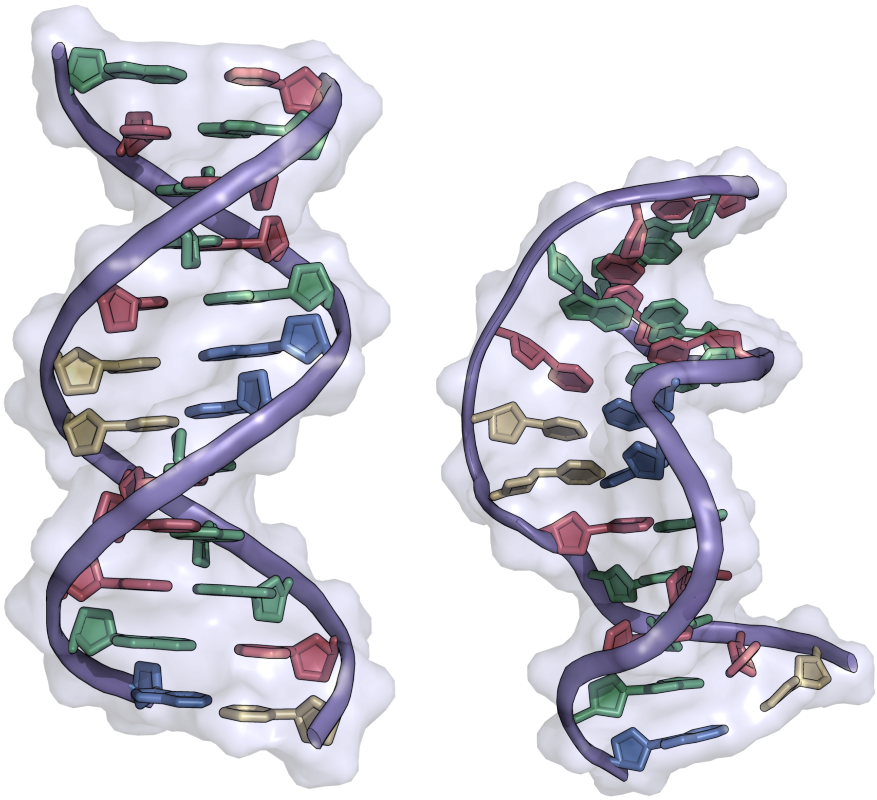


Figure A.4: Base pair steps and sequence of the DNA oligonucleotides. The DNA molecule included either two standard Thymine (T) or a central CPD damage (indicated by an enclosing gray box) opposite to the two adenine bases.



(a) B-DNA conformation of the simulated short DNA sequence. The simulations started from this structure are labeled as $\mathbf{TT_{BDNA}}$ for the undamaged sequence and $\mathbf{CPD_{BDNA}}$ for the damaged sequence.

(b) Deformed DNA structure by restraining all backbone atoms (except the damaged nucleotides) to the CPD-damaged DNA crystal structure in complex with the E. coli photolyase (pdb-entry: PDB:1TEZ [117]) The undamaged structure is labeled as $\mathbf{TT_{1TEZ}}$ and the damaged structure as $\mathbf{CPD_{1TEZ}}$.

Figure A.5: Structures used for MD simulations of the 1TEZ-system.

BIBLIOGRAPHY

[1] Abdelilah Aboussekhra et al. "Mammalian DNA Nucleotide Excision Repair Reconstituted with Purified Protein Components." In: *Cell* 80 (1995), pp. 859–868.

[2] Alan Grossfield. *An implementation of WHAM: the Weighted Histogram Analysis Method Version 2.0.9*, pp. 1–18.

[3] B. Alberts et al. *Molecular Biology of the Cell*. Molecular Biology of the Cell. Taylor & Francis, 2014. ISBN: 9780815344322.

[4] Robert a. Alberty. "Use of Legendre Transforms in Chemiccal Thermodynamics." In: *Pure Appl. Chem.* 73.8 (2001), pp. 1349–1380. ISSN: 00219614. DOI: 10.1351/pac200173081349.

[5] B. J. Alder and T. E. Wainwright. "Studies in Molecular Dynamics. I. General Method." In: *J. Chem. Phys.* 31.2 (1959), p. 459. ISSN: 00219606. DOI: 10.1063/1.1730376.

[6] Mp Allen. "Introduction to molecular dynamics simulation." In: *Comput. Soft Matter From Synth. Polym. to ...* 23 (2004), pp. 1–28. ISSN: 17412552. DOI: 10.1016/j.cplett.2006.06.020.

[7] Hans C Andersen. "Rattle: A velocity version of the shake algorithm for molecular dynamics calculations." In: *J. Comput. Phys.* 52 (1983), pp. 24–34. ISSN: 00219991. DOI: 10.1016/0021-9991(83)90014-1.

[8] User: AndrewKepert. *Tesselation of space using truncated octahedra*. [Online; accessed 2015-08-05]. Aug. 2012.

[9] Anjum Ansari. "Mean first passage time solution of the Smoluchowski equation: Application to relaxation dynamics in myoglobin." In: *J. Chem. Phys.* 112.5 (2000), pp. 2516–2522. ISSN: 00219606. DOI: 10.1063/1.480818.

[10] Hanne S. Antila and Emppu Salonen. "Polarizable force fields." In: *Methods Mol. Biol.* 924 (2013), pp. 215–241. ISSN: 10643745. DOI: 10.1007/978-1-62703-017-5-9.

[11] S. Arrhenius and Physics pamphlets. *Über die Dissociationswärme und den Einfluss der Temperatur auf den Dissociationsgrad der Elektrolyte*. Wilhelm Engelmann, 1889.

[12] Subhendu Sekhar Bag. *Course Name : Bio-Organic Chemistry*. [Online; accessed 2015-07-10].

[13] Christian Bartels. "Analyzing biased Monte Carlo and molecular dynamics simulations." In: *Chem. Phys. Lett.* 331.5-6 (2000), pp. 446–454. ISSN: 00092614. DOI: 10.1016/S0009-2614(00)01215-X.

[14] D P Batty and R D Wood. "Damage recognition in nucleotide excision repair of DNA." In: *Gene* 241.2 (Jan. 2000), pp. 193–204. ISSN: 0378-1119.

[15] B. Berg, A. Billoire, and D. Foerster. *Monte Carlo method for random surfaces*. berg1985, 1985. DOI: 10.1016/S0550-3213(85)80002-X.

[16] J.M. Berg et al. *Stryer Biochemie*. Spektrum Akademischer Verlag, 2009. ISBN: 9783827418005.

[17] Carol Bernstein et al. "DNA Damage , DNA Repair and Cancer." In: *New Res. Dir. DNA Repair* (2013), pp. 413–466. ISSN: 978-953-51-1114-6. DOI: 10.5772/53919.

[18] Harris Bernstein et al. "Cancer and Aging as Consequences of Un-repaired DNA Damage." In: *Nova Science Publishers* (2008).

[19] National Center for Biotechnology Information. *PubChem Compound Database; CID=9250*. National Center for Biotechnology Information.

[20] Dana Branzei and Marco Foiani. "Regulation of DNA repair throughout the cell cycle." In: *Nat. Rev. Mol. Cell Biol.* 9.4 (2008), pp. 297–308. ISSN: 1471-0072. DOI: 10.1038/nrm2351.

[21] Klaus Brettel and Martin Byrdin. "Reaction mechanisms of DNA photolyase." In: *Curr. Opin. Struct. Biol.* 20.6 (Dec. 2010), pp. 693–701. ISSN: 1879-033X. DOI: 10.1016/j.sbi.2010.07.003.

[22] Ab Britt. "Molecular genetics of DNA repair in higher plants." In: *Trends Plant Sci.* 4.1 (1999), pp. 20–25. ISSN: 1878-4372. DOI: 10.1016/s1360-1385(98)01355-7.

[23] S D Bruner, D P Norman, and G L Verdine. "Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA." In: *Nature* 403.6772 (Feb. 2000), pp. 859–66. ISSN: 0028-0836. DOI: 10.1038/35002510.

[24] Yuqin Cai et al. "Free Energy Profiles of Base Flipping in Intercalative Polycyclic Aromatic Hydrocarbon-Damaged DNA Duplexes: Energetic and Structural Relationships to Nucleotide Excision Repair Susceptibility." In: *Chem. Res. Toxicol.* 26.7 (July 2013), pp. 1115–1125. ISSN: 1520-5010. DOI: 10.1021/tx400156a.

[25] Thomas Carell and Robert Epple. "Repair of UV Light Induced DNA Lesions : A Comparative Study with Model Compounds." In: *Eur. J. Org. Chem.* (1998), pp. 1245–1258.

[26] T Carell et al. "The mechanism of action of DNA photolyases." In: *Curr. Opin. Chem. Biol.* 5.5 (Oct. 2001), pp. 491–8. ISSN: 1367-5931.

[27] D.A. Case et al. *Amber 12 reference manual*. 2012.

[28] D.A. Case et al. *Amber 14*. University of California, San Francisco. 2015.

[29] David Chandler. "Statistical mechanics of isomerization dynamics in liquids and the transition state approximation." In: *J. Chem. Phys.* 68.6 (1978), p. 2959. ISSN: 00219606. DOI: 10.1063/1.436049.

[30] CD Christ. "Basic ingredients of free energy calculations: a review." In: *J. Comput. ...* (2010). DOI: 10.1002/jcc.

[31] Kateřina Chválová, Viktor Brabec, and Jana Kašpárková. "Mechanism of the formation of DNA-protein cross-links by antitumor cisplatin." In: *Nucleic Acids Res.* 35.6 (2007), pp. 1812–1821. ISSN: 03051048. DOI: 10.1093/nar/gkm032.

[32] Glossary O F Class et al. "Glossary of class." In: *Pure Appl. Chem.* 67 (1995), pp. 1307–1375.

[33] Evangelos a. Coutsias, Chaok Seok, and Ken a. Dill. "Using quaternions to calculate RMSD." In: *J. Comput. Chem.* 25.15 (2004), pp. 1849–1857. ISSN: 01928651. DOI: 10.1002/jcc.20110.

[34] Ralf Dahm. "Friedrich Miescher and the discovery of DNA." In: *Dev. Biol.* 278.2 (2005), pp. 274–288. ISSN: 00121606. DOI: 10.1016/j.ydbio.2004.11.028.

[35] Bjørn Dalhus et al. "DNA base repair–recognition and initiation of catalysis." In: *FEMS Microbiol. Rev.* 33.6 (Nov. 2009), pp. 1044–78. ISSN: 1574-6976. DOI: 10.1111/j.1574-6976.2009.00188.x.

[36] Kelly L Damm and Heather a Carlson. "Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures." In: *Biophys. J.* 90.12 (2006), pp. 4558–4573. ISSN: 00063495. DOI: 10.1529/biophysj.105.066654.

[37] Tom Darden, Darrin York, and Lee Pedersen. "Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems." In: *J. Chem. Phys.* 98.12 (1993). DOI: 10.1063/1.464397.

[38] Andrew J Deans and Stephen C West. "DNA interstrand crosslink repair and cancer." In: *Nat. Rev. Cancer* 11.7 (2011), pp. 467–480. ISSN: 1474-175X. DOI: 10.1038/nrc3088.

[39] R Dickerson et al. "Definitions and nomenclature of nucleic acid structure parameters." In: *J. Mol. Biol.* 205.4 (1989), pp. 787–91. ISSN: 0022-2836.

[40] D. Djuranovic and B. Hartmann. "DNA Fine Structure and Dynamics in Crystals and in Solution: The Impact of BI/BII Backbone Conformations." In: *Biopolymers* 73.3 (2004), pp. 356–368. ISSN: 00063525. DOI: 10.1002/bip.10528.

[41] Lars-Oliver Essen. "Photolyases and cryptochromes: common mechanisms of DNA repair and light-driven signaling?" In: *Curr. Opin. Struct. Biol.* 16.1 (Feb. 2006), pp. 51–9. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2006.01.004.

[42] Fangqiang Zhu and Gerhard Hummer. "Convergence and Error Estimation in Free Energy Calculations Using the Weighted Histogram Analysis Method." In: *J. Comput. Chem.* (2011), pp. 453–465. DOI: 10.1002/jcc.21989.

[43] E Fermi, J Pasta, and S Ulam. "Studies of nonlinear problems." In: *LASL Rep. LA-1940* (1955).

[44] Alan M. Ferrenberg and Robert H. Swendsen. "New Monte Carlo technique for studying phase transitions." In: *Phys. Rev. Lett.* 61.23 (1988), pp. 2635–2638. ISSN: 00319007. DOI: 10.1103/PhysRevLett.61.2635.

[45] Alan M. Ferrenberg and Robert H. Swendsen. "Optimized Monte Carlo data analysis." In: *Phys. Rev. Lett.* 63.12 (1989), pp. 1195–1198. ISSN: 00319007. DOI: 10.1103/PhysRevLett.63.1195.

[46] Eric S Fischer et al. "The molecular basis of CRL4DDB2/CSA ubiquitin ligase architecture, targeting, and activation." In: *Cell* 147.5 (Nov. 2011), pp. 1024–39. ISSN: 1097-4172. DOI: 10.1016/j.cell.2011.10.035.

[47] James Fishburn et al. "Double-stranded DNA translocase activity of transcription factor TFIIH and the mechanism of RNA polymerase II open complex formation." In: *Proc. Natl. Acad. Sci.* 112.13 (2015), pp. 3961–3966. ISSN: 0027-8424. DOI: 10.1073/pnas.1417709112.

[48] Alex A Freitas and João Pedro De Magalhães. "A review and appraisal of the DNA damage theory of ageing." In: *Mutat. Res. - Rev. Mutat. Res.* 728.1-2 (2011), pp. 12–22. ISSN: 13835742. DOI: 10.1016/j.mrrev.2011.05.001.

[49] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Computational science series. Elsevier Science, 2001. ISBN: 9780080519982.

[50] E.C. Friedberg. *Correcting the Blueprint of Life: An Historical Account of the Discovery of DNA Repair Mechanisms*. Cold Spring Harbor Laboratory Press, 1997. ISBN: 9780879695071.

[51] EC Friedberg. "DNA damage and repair." In: *Nature* 421.January (2003).

[52] Errol C. Friedberg, G.C. Walker, and W. Siede. *DNA Repair and Mutagenesis*. ASM Press, 1995. ISBN: 9781555810887.

[53] Joshua I Friedman and James T Stivers. "Detection of damaged DNA bases by DNA glycosylase enzymes." In: *Biochemistry (Mosc.)* 49.24 (June 2010), pp. 4957–67. ISSN: 1520-4995. DOI: 10.1021/bi100593a.

[54] Monika Fuxreiter et al. "Role of Base Flipping in Specific Recognition of Damaged DNA by Repair Enzymes." In: *J. Mol. Biol.* 323.5 (Nov. 2002), pp. 823–834. ISSN: 00222836. DOI: 10.1016/S0022-2836(02)00999-3.

[55] Ruggero Gabbrielli. "A new counter-example to Kelvin's conjecture on minimal surfaces." In: *Philos. Mag. Lett.* March (2009), pp. 37–41. ISSN: 0950-0839. DOI: 10.1080/09500830903022651.

[56] Swapan K. Ghosh. "Nobel prize in chemistry 2013: Chemistry in cyberspace." In: *Curr. Sci.* 105.11 (2013), pp. 1455–1456. ISSN: 00113891.

[57] Stefano Gianni, Jakob Dogan, and Per Jemth. "Distinguishing induced fit from conformational selection." In: *Biophys. Chem.* 189 (2014), pp. 33–9. ISSN: 1873-4200. DOI: 10.1016/j.bpc.2014.03.003.

[58] J. B. Gibson et al. "Dynamics of radiation damage." In: *Phys. Rev.* 120.4 (1960), pp. 1229–1253. ISSN: 0031899X. DOI: 10.1103/PhysRev.120.1229.

[59] SS Gill, NA Anjum, and Ritu Gill. "DNA Damage and Repair in Plants under Ultraviolet and Ionizing Radiations." In: *Sci. World J.* (2014), pp. 5–7.

[60] Ohad Givaty and Yaakov Levy. "Protein sliding along DNA: dynamics and structural characterization." In: *J. Mol. Biol.* 385.4 (Jan. 2009), pp. 1087–97. ISSN: 1089-8638. DOI: 10.1016/j.jmb.2008.11.016.

[61] Andreas F Glas et al. "Crystal structure of the T(6-4)C lesion in complex with a (6-4) DNA photolyase and repair of UV-induced (6-4) and Dewar photolesions." In: *Chemistry - A European Journal* 15.40 (Oct. 2009), pp. 10387–96. ISSN: 1521-3765. DOI: 10.1002/chem.200901004.

[62] Andreas F Glas et al. "DNA (6-4) photolyases reduce Dewar isomers for isomerization into (6-4) lesions." In: *J. Am. Chem. Soc.* 132.10 (Mar. 2010), pp. 3254–5. ISSN: 1520-5126. DOI: 10.1021/ja910917f.

[63] User: Grimlock. *Scheme of periodic tiles for space*. [Online; accessed 2015-08-05]. Aug. 2007.

[64] Sami N. Guzder et al. "Affinity of yeast nucleotide excision repair factor 2, consisting of the Rad4 and Rad23 proteins, for ultraviolet damaged DNA." In: *J. Biol. Chem.* 273.47 (1998), pp. 31541–31546. ISSN: 00219258. DOI: 10.1074/jbc.273.47.31541.

[65] Peter Hänggi and Michal Borkovec. "Reaction-rate theory: fifty years after Kramers." In: *Rev. Mod. Phys.* 62.2 (1990), pp. 251–341. ISSN: 0034-6861. DOI: 10.1103/RevModPhys.62.251.

[66] Brian D Harfe and S Jinks-Robertson. "DNA mismatch repair and genetic instability." In: *Annu. Rev. Genet.* 34 (2000), pp. 359–399. ISSN: 0066-4197. DOI: 10.1146/annurev.genet.34.1.359.

[67] Alan Herbert and Alexander Rich. "The Biology of Left-handed Z-DNA." In: *J. Biol. Chem.* 26 (1996).

[68] Berk Hess. "Determining the shear viscosity of model liquids from molecular dynamics simulations." In: *J. Chem. Phys.* 116.1 (2002), pp. 209–217. ISSN: 00219606. DOI: 10.1063/1.1421362.

[69] Jan H J Hoeijmakers. "DNA Damage, Aging, and Cancer." In: *The New England Journal of Medicine* (2009), pp. 1475–1485.

[70] David S. Hsu et al. "Putative human blue-light photoreceptors hCRY1 and hCRY2 are flavoproteins." In: *Biochemistry (Mosc.)* 35.44 (1996), pp. 13871–13877. ISSN: 00062960. DOI: 10.1021/bi962209o.

[71] Niu Huang and Alexander D MacKerell. "Atomistic view of base flipping in DNA." In: *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 362.1820 (July 2004), pp. 1439–60. ISSN: 1364-503X. DOI: 10.1098/rsta.2004.1383.

[72] Yaling Huang and Lei Li. "DNA crosslinking damage and cancer - a tale of friend and foe." In: *Transl. Cancer Res.* 2.3 (2013), pp. 144–154. ISSN: 2218-676X. DOI: 10.3978/j.issn.2218-676X.2013.03.01. arXiv: NIHMS150003.

[73] Gerhard Hummer. "Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations." In: *New J. Phys.* 7 (2005). ISSN: 13672630. DOI: 10.1088/1367-2630/7/1/034.

[74] J. D. Hunter. "Matplotlib: A 2D graphic environment." In: *Comput. Sci. Eng.* 9.3 (2007), pp. 90–95. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.55.

[75] I Husain, J Griffith, and a Sancar. "Thymine dimers bend DNA." In: *Proc. Natl. Acad. Sci. U. S. A.* 85.8 (Apr. 1988), pp. 2558–62. ISSN: 0027-8424.

[76] I Husian and Aziz Sancar. "Binding of E. coli DNA photolyase to a defined substrate containing a single T<> T dimer." In: *Nucleic Acids Res.* 15.3 (1987).

[77] Ravi R Iyer et al. "DNA mismatch repair: functions and mechanisms." In: *Chem. Rev.* 106.2 (2006), pp. 302–23. ISSN: 0009-2665. DOI: 10.1021/cr0404794.

[78] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. [Online; accessed 2015-09-21]. 2001–.

[79] William L Jorgensen et al. "Comparison of simple potential functions for simulating liquid water." In: *The Journal of Chemical Physics* 79.2 (1983).

[80] In Suk Joung and Thomas E Cheatham. "Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations." In: *J. Phys. Chem. B* 112.30 (July 2008), pp. 9020–9041. ISSN: 1520-6106. DOI: 10.1021/jp8001614.

[81] IS Joung and TE Cheatham III. "Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters." In: *J. Phys. Chem. B* (2009), pp. 13279–13290.

[82] W. Kabsch. "A solution for the best rotation to relate two sets of vectors." In: *Acta Crystallogr. Sect. A* 32.5 (Sept. 1976), pp. 922–923. ISSN: 0567-7394. DOI: 10.1107/S0567739476001873. arXiv: 05677394.

[83] Mahmut Kara and Martin Zacharias. "Influence of 8-oxoguanosine on the fine structure of DNA studied with biasing-potential replica exchange simulations." In: *Biophys. J.* 104.5 (2013), pp. 1089–1097. ISSN: 00063495. DOI: 10.1016/j.bpj.2013.01.032.

[84] Martin Karplus and J Andrew Mccammon. "Molecular dynamics simulations of biomolecules." In: *Nat. Struct. Biol.* 9.9 (2002).

[85] 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." In: *Photochem. Photobiol.* 62.1 (1995), pp. 44–50. ISSN: 1751-1097. DOI: 10.1111/j.1751-1097.1995.tb05236.x.

[86] T. Kim. "Mechanism of ATP-Dependent Promoter Melting by Transcription Factor IIH." In: *Science (80-.).* 288.5470 (May 2000), pp. 1418–1421. ISSN: 00368075. DOI: 10.1126/science.288.5470.1418.

[87] Stephan Kiontke et al. “Crystal structures of an archaeal class II DNA photolyase and its complex with UV-damaged duplex DNA.” In: *The EMBO journal* 30.21 (Nov. 2011), pp. 4437–49. ISSN: 1460-2075. DOI: 10.1038/emboj.2011.313.

[88] John G. Kirkwood. “Statistical mechanics of fluid mixtures.” In: *J. Chem. Phys.* 3.1935 (1935), pp. 300–313. ISSN: 00219606. DOI: 10.1063/1.1749657.

[89] Alexander Knips and Martin Zacharias. “Influence of a cis,syn-cyclobutane pyrimidine dimer damage on DNA conformation studied by molecular dynamics simulations.” In: *Biopolymers* 103.4 (Nov. 2014), pp. 215–222. ISSN: 0006-3525. DOI: 10.1002/bip.22586.

[90] Jiri Kolafa and John W. Perram. “Cutoff Errors in the Ewald Summation Formulae for Point Charge Systems.” In: *Mol. Simul.* 9.5 (1992), pp. 351–368. ISSN: 0892-7022. DOI: 10.1080/08927029208049126.

[91] R D Kornberg and Y Lorch. “Twenty-five years of the nucleosome, fundmamental particle of the eukaryotic chromosome.” In: *Cell* 98 (1999), pp. 285–294.

[92] Shankar Kumar et al. “THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method.” In: *J. Comput. Chem.* 13.8 (1992), pp. 1011–1021. ISSN: 1096-987X. DOI: 10.1002/jcc.540130812.

[93] Jochen Kuper and Caroline Kisker. “Damage recognition in nucleotide excision DNA repair.” In: *Curr. Opin. Struct. Biol.* 22.1 (Mar. 2012), pp. 88–93. ISSN: 1879-033X. DOI: 10.1016/j.sbi.2011.12.002.

[94] Andres a. Larrea, Scott a. Lujan, and Thomas a. Kunkel. “SnapShot: DNA Mismatch Repair.” In: *Cell* 141.4 (2010). ISSN: 00928674. DOI: 10.1016/j.cell.2010.05.002.

[95] R Lavery et al. “Conformational analysis of nucleic acids revisited: Curves+.” In: *Nucleic Acids Res.* 37.17 (Sept. 2009), pp. 5917–29. ISSN: 1362-4962. DOI: 10.1093/nar/gkp608.

[96] J H Lee, S H Bae, and B S Choi. “The Dewar photoproduct of thymidylyl(3’–>5’)- thymidine (Dewar product) exhibits mutagenic behavior in accordance with the "A rule".” In: *Proc. Natl. Acad. Sci. U. S. A.* 97.9 (Apr. 2000), pp. 4591–6. ISSN: 0027-8424. DOI: 10.1073/pnas.080057097.

[97] JH Lee, YJ Choi, and BS Choi. “Solution structure of the DNA decamer duplex containing a 3’-T-T base pair of the cis-syn cyclobutane pyrimidine dimer: implication for the mutagenic property of.” In: *Nucleic Acids Res.* 28.8 (2000), pp. 1794–1801.

[98] JH Lee, GS Hwang, and BS Choi. "Solution structure of a DNA decamer duplex containing the stable 3'-T-G base pair of the pyrimidine (6-4) pyrimidone photoproduct : Implications for." In: *Proc. Natl. Acad. Sci. U. S. A.* 96.June (1999), pp. 6632–6636.

[99] Joon-Hwa Lee et al. "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." In: *Nucleic Acids Res.* 32.8 (Jan. 2004), pp. 2474–81. ISSN: 1362-4962. DOI: 10.1093/nar/gkh568.

[100] S.W. de Leeuw, C.P. Williams, and B. Smit. "Evidence of phase separation in mixtures of Lennard-Jones and Stockmayer fluids." In: *Mol. Phys.* 65.5 (1988), pp. 1269–1272. ISSN: 0026-8976. DOI: 10.1080/00268978800101771.

[101] M Levitt. "A simplified representation of protein conformations for rapid simulation of protein folding." In: *J. Mol. Biol.* 104.1 (1976), pp. 59–107. ISSN: 00222836. DOI: 10.1016/0022-2836(76)90004-8.

[102] M Levitt and A Warshel. "Computer simulation of protein folding." In: *Nature* 253.5494 (1975), pp. 694–698. ISSN: 0028-0836. DOI: 10.1038/253694a0.

[103] Michael Levitt. "Molecular dynamics of native protein. II. Analysis and nature of motion." In: *J. Mol. Biol.* 168.3 (1983), pp. 621–57. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(83)80304-0.

[104] Y F Li, S T Kim, and A Sancar. "Evidence for lack of DNA photoreactivating enzyme in humans." In: *Proc. Natl. Acad. Sci. U. S. A.* 90.10 (1993), pp. 4389–4393. ISSN: 0027-8424. DOI: 10.1073/pnas.90.10.4389.

[105] T Lindahl. "Instability and decay of the primary structure of DNA." In: *Nature* 362.6422 (1993), pp. 709–715. ISSN: 0028-0836. DOI: 10.1038/362709a0.

[106] Harvey Lodish et al. "Molecular Cell Biology (4th edition)." In: *Free. Co., New York, NY, 2000* 29.3 (2000), pp. 126–128. ISSN: 14708175. DOI: 10.1016/S1470-8175(01)00023-6.

[107] Xiang J. Lu and Wilma K. Olson. "3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures." In: *Nucleic Acids Res.* 31.17 (2003), pp. 5108–5121. ISSN: 03051048. DOI: 10.1093/nar/gkg680.

[108] Xiang-Jun Lu and Wilma K Olson. "3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures." In: *Nat. Protoc.* 3.7 (2008), pp. 1213–1227. ISSN: 1754-2189. DOI: 10.1038/nprot.2008.104.

[109] K Luger et al. "Crystal structure of the nucleosome core particle at 2.8 A resolution." In: *Nature* 389.6648 (1997), pp. 251–260. ISSN: 0028-0836. DOI: 10.1038/38444.

[110] Alexander D. Mackerell. "Empirical force fields for biological macromolecules: Overview and issues." In: *J. Comput. Chem.* 25.13 (2004), pp. 1584–1604. ISSN: 01928651. DOI: 10.1002/jcc.20082.

[111] User: Madprime. *Chemical structure of DNA*. [Online; accessed 2015-06-05]. Mar. 2013.

[112] F Masson and T Laino. "Computational study of thymine dimer radical anion splitting in the self-repair process of duplex DNA." In: *J. Am. Chem. Soc.* 7863.10 (2008), pp. 7095–7104.

[113] K McAteer et al. "Solution-state structure of a DNA dodecamer duplex containing a Cis-syn thymine cyclobutane dimer, the major UV photoproduct of DNA." In: *J. Mol. Biol.* 282.5 (Oct. 1998), pp. 1013–32. ISSN: 0022-2836. DOI: 10.1006/jmbi.1998.2062.

[114] Wes McKinney. "Data Structures for Statistical Computing in Python." In: *Proc. 9th Python Sci. Conf.* 1697900.Scipy (2010), pp. 51–56.

[115] a. D. McNaught. "Nomenclature of carbohydrates. Recommendations 1996." In: *Carbohydr. Res.* 297.1 (1997), pp. 1–92. ISSN: 00086215. DOI: 10.1016/S0008-6215(97)83449-0.

[116] Ewan M. McNeil and David W. Melton. "DNA repair endonuclease ERCC1-XPF as a novel therapeutic target to overcome chemoresistance in cancer therapy." In: *Nucleic Acids Res.* 40.20 (2012), pp. 9990–10004. ISSN: 03051048. DOI: 10.1093/nar/gks818.

[117] Alexandra Mees et al. "Crystal structure of a photolyase bound to a CPD-like DNA lesion after in situ repair." In: *Science* 306.5702 (Dec. 2004), pp. 1789–93. ISSN: 1095-9203. DOI: 10.1126/science.1101598.

[118] M.K Memon, R.W Hockney, and S.K Mitra. "Molecular Dynamics with Constraints." In: *J. Comput. Phys.* 43 (1981), pp. 345–356. ISSN: 00219991. DOI: 10.1016/0021-9991(81)90127-3.

[119] Karol Miaskiewicz et al. "Computational simulations of DNA distortions by a cis, syn-cyclobutane thymine dimer lesion." In: *J. Am. Chem. Soc.* 118.38 (1996), pp. 9156–9163.

[120] Maria Mills and Ioan Andricioaei. "An experimentally guided umbrella sampling protocol for biomolecules." In: *J. Chem. Phys.* 129.11 (2008), pp. 1–13. ISSN: 00219606. DOI: 10.1063/1.2976440.

[121] David L Mitchell and Rodney S Nairn. "The biology of the (6–4) photoproduct." In: *Photochem. Photobiol.* 49.6 (1989), pp. 805–819. ISSN: 1751-1097. DOI: 10.1111/j.1751-1097.1989.tb05578.x.

[122] Y. Mitsui et al. "Physical and Enzymatic Studies on Poly d(I-C) Poly d(I-C), an Unusual Double-helical DNA." In: *Nature* (1970).

[123] Ayori Mitsutake, Yuji Sugita, and Yuko Okamoto. "Generalized-ensemble algorithms for molecular simulations of biopolymers." In: *Biopolym. - Pept. Sci. Sect.* 60.2 (2001), pp. 96–123. ISSN: 00063525. DOI: 10.1002/1097-0282(2001)60:2<96::AID-BIP1007>3.0.CO;2-F. arXiv: 0012021 [cond-mat].

[124] J K Moore and J E Haber. "Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae." In: *Mol. Cell. Biol.* 16.5 (1996), pp. 2164–2173. ISSN: 0270-7306.

[125] G. P. Moss. "Basic terminology of stereochemistry (IUPAC Recommendations 1996)." In: *Pure Appl. Chem.* 68.12 (1996), pp. 2193–2222. ISSN: 0033-4545. DOI: 10.1351/pac199668122193.

[126] D. Mu et al. "Reconstitution of human DNA repair excision nuclease in a highly defined system." In: *J. Biol. Chem.* 270.6 (1995), pp. 2415–2418. ISSN: 00219258. DOI: 10.1074/jbc.270.6.2415.

[127] Markus Müller and Thomas Carell. "Structural biology of DNA photolyases and cryptochromes." In: *Curr. Opin. Struct. Biol.* 19.3 (June 2009), pp. 277–85. ISSN: 1879-033X. DOI: 10.1016/j.sbi.2009.05.003.

[128] Lauren L O'Neil, Alan Grossfield, and Olaf Wiest. "Base flipping of the thymine dimer in duplex DNA." In: *J. Phys. Chem. B* 111.40 (Oct. 2007), pp. 11843–9. ISSN: 1520-6106. DOI: 10.1021/jp074043e.

[129] L.L. Lauren L O'Neil, Olaf Wiest, and L.L. O'Nei. "Structures and energetics of base flipping of the thymine dimer depend on DNA sequence." In: *J. Phys. Chem. B* 112.13 (Apr. 2008), pp. 4113–22. ISSN: 1520-6106. DOI: 10.1021/jp7102935.

[130] R Osman, M Fuxreiter, and N Luo. "Specificity of damage recognition and catalysis of DNA repair." In: *Comput. Chem.* 24.3-4 (2000), pp. 331–339. ISSN: 00978485. DOI: 10.1016/S0097-8485(99)00073-X.

[131] C O Pabo and R T Sauer. "Protein-DNA Recognition." In: *Annu. Rev. Biochem.* 53.1 (1984), pp. 293–321. DOI: 10.1146/annurev.bi.53.070184.001453.

[132] Q Pang and J B Hays. "UV-B-Inducible and Temperature-Sensitive Photoreactivation of Cyclobutane Pyrimidine Dimers in Arabidopsis thaliana." In: *Plant Physiol.* 95.2 (1991), pp. 536–543. ISSN: 0032-0889. DOI: 10.1104/pp.95.2.536.

[133] HaJeung Park et al. "Crystal structure of a DNA decamer containing a cis-syn thymine dimer." In: *Proc. Natl. Acad. Sci. U. S. A.* 99.25 (2002), pp. 15965–15970.

[134] Jared B Parker et al. "Enzymatic capture of an extrahelical thymine in the search for uracil in DNA." In: *Nature* 449.7161 (Sept. 2007), pp. 433–7. ISSN: 1476-4687. DOI: 10.1038/nature06131.

[135] Alberto Pérez et al. "Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers." In: *Biophys. J.* 92.11 (June 2007), pp. 3817–29. ISSN: 0006-3495. DOI: 10.1529/biophysj.106.097782.

[136] Fritz M. Pohl and Thomas M. Jovin. "Salt-induced co-operative conformational change of a synthetic DNA: Equilibrium and kinetic studies with poly(dG-dC)." In: *J. Mol. Biol.* 67.3 (1972), pp. 375–396. ISSN: 00222836. DOI: 10.1016/0022-2836(72)90457-3.

[137] Andrew Pohorille, Christopher Jarzynski, and Christophe Chipot. "Good practices in free-energy calculations." In: *J. Phys. Chem. B* 114.32 (2010), pp. 10235–10253. ISSN: 15206106. DOI: 10.1021/jp102971x.

[138] Sinden RR Potaman VN. *DNA: Alternative Conformations and Biology*. Madame Curie Bioscience Database [Internet]. Austin (TX): Landes Bioscience. [Online; accessed 2015-09-25].

[139] A Rahman. "Correlations in the Motion Atoms in Liquid Argon." In: *Phys. Rev.* 136.2 A (1964), A 405 –A 411. ISSN: 0031-899X. DOI: 10.1103/PhysRev.136.A405.

[140] Rajesh P Rastogi et al. "Molecular mechanisms of ultraviolet radiation-induced DNA damage and repair." In: *J. Nucleic Acids* 2010 (Jan. 2010), p. 592980. ISSN: 2090-021X. DOI: 10.4061/2010/592980.

[141] K M Reinisch et al. "The crystal structure of HaeIII methyltransferase convalently complexed to DNA: an extrahelical cytosine and rearranged base pairing." In: *Cell* 82.1 (1995), pp. 143–153. ISSN: 0092-8674. DOI: 10.1016/0092-8674(95)90060-8.

[142] Benoît Roux. *The calculation of the potential of mean force using computer simulations*. 1995. DOI: 10.1016/0010-4655(95)00053-I.

[143] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes." In: *J. Comput. Phys.* 23 (1977), pp. 327–341. ISSN: 00219991. DOI: 10.1016/0021-9991(77)90098-5.

[144] E . R . Smith S . W . de Leeuw , J . W . Perram. "Simulation of Electrostatic Systems in Periodic Boundary Conditions. I. Lattice Sums and Dielectric Constants." In: *Proc. R. Soc. Lond. A* 373.1752 (1980), pp. 27–56. ISSN: 1364-5021. DOI: 10.1098/rspa.1980.0135.

[145] E . R . Smith S . W . de Leeuw , J . W . Perram. "Simulation of Electrostatic Systems in Periodic Boundary Conditions . II . Equivalence of Boundary Conditions." In: *Proc. R. Soc. Lond. A* 373.1752 (1980), pp. 57–66.

[146] Aziz Sancar. "Structure and function of DNA photolyase and cryptochrome blue-light photoreceptors." In: *Chem. Rev.* 103.6 (2003), pp. 2203–2238.

[147] B. Schneider, S. Neidle, and H. M. Berman. "Conformations of the sugar-phosphate backbone in helical DNA crystal structures." In: *Biopolymers* 42.1 (1997), pp. 113–124. ISSN: 00063525. DOI: 10.1002/(SICI)1097-0282(199707)42:1<113::AID-BIP10>3.0.CO;2-O.

[148] Andrea Scrima et al. "Structural basis of UV DNA-damage recognition by the DDB1-DDB2 complex." In: *Cell* 135.7 (Dec. 2008), pp. 1213–23. ISSN: 1097-4172. DOI: 10.1016/j.cell.2008.10.045.

[149] Rajeshwar P Sinha and Donat P Häder. "UV-induced DNA damage and repair: a review." In: *Photochem. Photobiol. Sci.* 1.4 (2002), pp. 225–236. ISSN: 1474905X. DOI: 10.1039/b201230h.

[150] Geir Slupphaug et al. "A nucletide-flipping mechanism from the structure of human uracil-DNA glycosylase bound to DNA." In: *Nature* 384 (1996), pp. 356–358.

[151] M. von Smoluchowski. "Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen." In: *Ann. Phys.* 326.14 (1906), pp. 756–780. ISSN: 00033804. DOI: 10.1002/andp.19063261405.

[152] Kun Song et al. "An Improved Reaction Coordinate for Nucleic Acid Base Flipping Studies." In: *J. Chem. Theory Comput.* 5.11 (Nov. 2009), pp. 3105–3113. ISSN: 1549-9618. DOI: 10.1021/ct9001575.

[153] TI Spector, Thomas E. Cheatham, and Peter A. Kollman. "Unrestrained molecular dynamics of photodamaged DNA in aqueous solution." In: *J. Am. Chem. Soc.* 7863.10 (1997), pp. 7095–7104.

[154] M. Ashley Spies and Richard L. Schowen. "The trapping of a spontaneously "flipped-out" base from double helical nucleic acids by host-guest complexation with β-cyclodextrin: The intrinsic base-flipping rate constant for DNA and RNA." In: *J. Am. Chem. Soc.* 124.47 (2002), pp. 14049–14053. ISSN: 00027863. DOI: 10.1021/ja012272n.

[155] Justin Spiriti et al. "DNA Bending through Large Angles Is Aided by Ionic Screening." In: *J. Chem. Theory Comput.* 8.6 (June 2012), pp. 2145–2156. ISSN: 1549-9618. DOI: 10.1021/ct300177r.

[156] User: Sponk. *Comparison of a single-stranded RNA and a double-stranded DNA with their corresponding nucleobases*. [Online; accessed 2015-08-05]. Mar. 2010.

[157] Claudia Steffen et al. "Unorthodox Uses of Bennett's Acceptance Ratio Method." In: *J. Comput. Chem.* 31.16 (2010), pp. 2967–2970. ISSN: 1096-987X. DOI: 10.1002/jcc.

[158] E Stofer and R Lavery. "Measuring the geometry of DNA grooves." In: *Biopolymers* 34.3 (1994), pp. 337–346. ISSN: 0006-3525. DOI: 10.1002/bip.360340305.

[159] Kaoru Sugasawa et al. "A molecular mechanism for DNA damage recognition by the xeroderma pigmentosum group C protein complex." In: *DNA Repair (Amst).* 1 (2002), pp. 95–107.

[160] DL Svoboda et al. "Effect of sequence, adduct type, and opposing lesions on the binding and repair of ultraviolet photodamage by DNA photolyase and (A) BC excinuclease." In: *J. Biol. Chem.* 268.14 (1993), pp. 10694–10700.

[161] Robert H. Swendsen and Jian Sheng Wang. "Replica Monte Carlo simulation of spin-glasses." In: *Phys. Rev. Lett.* 57.21 (1986), pp. 2607–2609. ISSN: 00319007. DOI: 10.1103/PhysRevLett.57.2607.

[162] A Szabo, K Schulten, and Z Schulten. "First passage time approach to diffusion controlled reactions." In: *J. Chem. ...* 72.1980 (1980), pp. 4350–7. ISSN: 00219606. DOI: 10.1063/1.439715.

[163] Yuichi Takeuchi et al. "The Photorepair and Photoisomerization of DNA Lesions in Etiolated Cucumber Cotyledons after Irradiation by UV-B Depends on Wavelength." In: *Plant Cell Physiol.* 39.7 (1998), pp. 745–750. ISSN: 0032-0781. DOI: 10.1093/oxfordjournals.pcp.a029429.

[164] Thomas J. Thamann et al. "The high salt form of poly(dG-dC)·poly(dG-dC) is left-handed Z-DNA: Raman spectra of crystals and solutions." In: *Nucleic Acids Res.* 9.20 (1981), pp. 5443–5458. ISSN: 03051048. DOI: 10.1093/nar/9.20.5443.

[165] G.M. Torrie and J.P. Valleau. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling." In: *J. Comput. Phys.* 23.2 (1977), pp. 187–199. ISSN: 00219991. DOI: 10.1016/0021-9991(77)90121-8.

[166] a a Travers. "DNA conformation and protein binding." In: *Annu. Rev. Biochem.* 58 (1989), pp. 427–452. ISSN: 00664154. DOI: 10.1146/annurev.biochem.58.1.427.

[167] N Tuteja et al. "Molecular mechanisms of DNA damage and repair: progress in plants." In: *Crit. Rev. Biochem. Mol. Biol.* 36.4 (2001), pp. 337–397. ISSN: 1040-9238. DOI: 10.1080/20014091074219.

[168] Stéfan Van Der Walt, S. Chris Colbert, and Gaël Varoquaux. "The NumPy array: A structure for efficient numerical computation." In: *Comput. Sci. Eng.* 13.2 (2011), pp. 22–30. ISSN: 15219615. DOI: 10.1109/MCSE.2011.37. arXiv: 1102.1523.

[169] Wilfred F. Van Gunsteren et al. *Biomolecular modeling: Goals, problems, perspectives*. 2006. DOI: 10.1002/anie.200502655.

[170] Péter Várnai and Richard Lavery. "Base flipping in DNA: pathways and energetics studied with molecular dynamic simulations." In: *J. Am. Chem. Soc.* 124.25 (June 2002), pp. 7272–3. ISSN: 0002-7863.

[171] Loup Verlet. "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules." In: *Phys. Rev.* 159.1 (1967), pp. 98–103.

[172] A H Wang et al. "Molecular structure of a left-handed double helical DNA fragment at atomic resolution." In: *Nature* 282.5740 (1979), pp. 680–686. ISSN: 0028-0836. DOI: 10.1038/282680a0.

[173] Cheng I. Wang and John S. Taylor. "In vitro evidence that UV-induced frameshift and substitution mutations at T tracts are the result of misalignment-mediated replication past a specific thymine dimer." In: *Biochemistry (Mosc.)* 31.14 (1992), pp. 3671–3681. DOI: 10.1021/bi00129a016. eprint: http://dx.doi.org/10.1021/bi00129a016.

[174] CI Wang and JS Taylor. "Site-specific effect of thymine dimer formation on dAn. dTn tract bending and its biological implications." In: *Proc. Natl. Acad. Sci. U. S. A.* 88.October (1991), pp. 9072–9076.

[175] Ariel Warshel. "Bicycle-pedal model for the first step in the vision process." In: *Group* 260.5552 (1976), pp. 619–21. ISSN: 0028-0836. DOI: 10.1038/260170a0.

[176] F. H. C. Watson, J. D.; Crick. "Molecular Structure of Nucleic Acids." In: *Nature* 171 (1953), pp. 737–738. ISSN: 0028-0836. DOI: 10.1038/171737a0.

[177] G H Weiss. "First passage time problems in chemical physics." In: *Adv. Chem. Phys.* XIII.August 1965 (1967), pp. 1–18.

[178] Peter H Berens William C. Swope Hans C. Andersen 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." In: *J. Chem. Phys.* 76.1 (1982), pp. 637–649. ISSN: 0021-9606. DOI: 10.1063/1.442716.

[179] Thomas J Wilson et al. "Evidence from thermodynamics that DNA photolyase recognizes a solvent-exposed CPD lesion." In: *The journal of physical chemistry B* 115.46 (Nov. 2011), pp. 13746–54. ISSN: 1520-5207. DOI: 10.1021/jp208129a.

[180] F K Winkler. "DNA totally flipped-out by methylase." In: *Structure* 2.2 (1994), pp. 79–83. ISSN: 09692126. DOI: 10.1016/S0969-2126(00)00009-5.

[181] Peter Yakovchuk, Ekaterina Protozanova, and Maxim D. Frank-Kamenetskii. "Base-stacking and base-pairing contributions into thermal stability of the DNA double helix." In: *Nucleic Acids Res.* 34.2 (2006), pp. 564–574. ISSN: 03051048. DOI: 10.1093/nar/gkj454.

[182] Hideshi Yokoyama, Ryuta Mizutani, and Yoshinori Satow. "Structure of a double-stranded DNA (6-4) photoproduct in complex with the 64M-5 antibody Fab." In: *Acta Crystallogr. D Biol. Crystallogr.* 69.Pt 4 (Apr. 2013), pp. 504–12. ISSN: 1399-0047. DOI: 10.1107/S0907444912050007.

[183] Richard Wheeler - User: Zephyris. *From left to right, the structures of A-, B- and Z-DNA.* [Online; accessed 2015-08-05]. May 2012.

[184] Marie Zgarbová et al. "Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles." In: *J. Chem. Theory Comput.* 7.9 (Sept. 2011), pp. 2886–2902. ISSN: 1549-9618. DOI: 10.1021/ct200162x.

[185] D O Zharkov. "Base excision DNA repair." In: *Cell. Mol. Life Sci.* 65.10 (May 2008), pp. 1544–65. ISSN: 1420-682X. DOI: 10.1007/s00018-008-7543-2.

[186] Han Zheng et al. "Base flipping free energy profiles for damaged and undamaged DNA." In: *Chem. Res. Toxicol.* 23.12 (Dec. 2010), pp. 1868–70. ISSN: 1520-5010. DOI: 10.1021/tx1003613.

[187] Ruhong Zhou. "Replica exchange molecular dynamics method for protein folding simulation." In: *Methods Mol. Biol.* 350.November (Jan. 2007), pp. 205–23. ISSN: 1064-3745.

[188] Fangqiang Zhu and Gerhard Hummer. "Convergence and error estimation in free energy calculations using the weighted histogram analysis method." In: *J Comput Chem.* 29.4 (2012), pp. 997–1003. ISSN: 15378276. DOI: 10.1016/j.biotechadv.2011.08.021.Secreted. arXiv: NIHMS150003.

[189] Fangqiang Zhu and Gerhard Hummer. "Theory and Simulation of Ion Conduction in the Pentamerice GLIC Channel." In: *J. Chem. Theory Comput.* bart (2011).

[190] R Zwanzig. "Diffusion in a rough potential." In: *Proc. Natl. Acad. Sci. U. S. A.* 85.7 (1988), pp. 2029–2030. ISSN: 0027-8424. DOI: 10.1073/pnas.85.7.2029.

ACKNOWLEDGEMENTS

Finally, I would like to thank all the people who supported me throughout my PhD:

- Univ.-Prof. Dr. Martin Zacharias for being such a magnificent mentor to me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on topic of research as well as on my career have been priceless.
- Dr. Nadine Schwierz and Christina Frost for suggestions and discussion about the calculation of Mean First Passage Times (MFPTs) by means of Umbrella Sampling (US) simulation data.
- Giuseppe La Rosa for being a wonderful person to collaboratively work together on DNA repair in general. I thoroughly enjoyed our discussions, work trips, and meanderings about not-physics related topics.
- Florian Kandzia, Florian Häse, Fabian Zeller, Rainer Bomblies, Korbinian Liebl, and Max Liebkies for reading over my manuscript.
- Sonja Ortner for helping me with administrative problems.
- The whole group of T38 for a wonderful time in a fantastic atmosphere.
- My family for their special support. Words cannot express how grateful I am to my mother, and father for all sacrifices that you've made on my behalf.
- Lukas Knips for being supportive on so many levels, professionally and non-professionally.
- My friends who supported me in writing, and encouraged me to strive towards my goal.

At the end I have to thank the SFB 749 (Dynamics and Intermediates of Molecular Transformations) for providing the funding for the research of my PhD thesis.