Molecular Modelling Studies of DNA Damage Recognition

Supat Jiranusornkul,

B.Pharm. (Hons.), M.Pharm. (Pharm. Chem.)

Molecular Recognition Group

School of Pharmacy

University of Nottingham

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy, April 2008

The Music of Life (in the key of DNA) [1].

Abstract

How DNA repair proteins search and recognise the rare sites of damage from the massive numbers of normal DNA remains poorly understood. FapydG (2,6-diamino-4-hydroxy-5-formamidopyrimidine) is one of the most prevalent guanine derived lesions involving opening of the imidazole ring. It is typically repaired by formamidopyrimidine-DNA glycosylase (Fpg) as an initial step in base excision repair; if not repaired, the lesion generates a G:C → T:A transversion. Unfortunately, studies on the recognition of FapydG have been hindered by difficulties to synthesise and incorporate the FapydG residue into a DNA duplex. Crystal structures of Fpg-DNA complexes have demonstrated three common recognition events: the protein specifically binding to the extrahelical lesion, bending DNA centred on the damaged base, and flipping the damage into the pocket. Thus, molecular modelling and dynamics simulation have been used to gather dynamical information of those recognition events for damaged and undamaged DNA. The simulations were initially performed when FapydG or G occurs in several dodecamer B-DNA sequences in aqueous solution, then inside the lesion-recognition pocket of Fpg, and during the flipping pathway from the helical stack to an extrahelical position.

The influence of the damage on DNA stability and flexibility was first investigated. Energetic analysis revealed that damage to DNA does appear to destabilise the duplex. DNA curvature analysis and a novel combined method of the principal component analysis (PCA) and the Mahalanobis distance (D_M) indicated that damaged DNA can adopt the observed protein-bound conformation

with lower energetic penalties than its normal counterpart. Results of these studies have provided the validation of DNA bending enhancement by the FapydG lesion. It also suggested that intrinsic DNA bending could be a principal element of how the repair protein locates the lesion from vast expanse of normal bases.

Considering the specific recognition of FapydG by Fpg, the αF-β9 loop of the Fpg enzyme may function as a gatekeeping to accommodate the lesion while denying the normal base. Remarkably fluctuating movement of the flipped G residue and the αF-β9 loop is due to the formation of the non-specific Fpg/G complex with a lower binding energy by 8.4 *kcal/mol* compared to the specific Fpg/FapydG complex. Free-energy profiles for both damaged and undamaged base flipping were generated from the umbrella sampling simulations and the Weight Histogram Analysis Method (WHAM). An energy barrier for flipping the damage out from the helix is 2.7 *kcal/mol* higher than its equivalent G and the lesion is highly stabilised inside the pocket. In contrast, G flipping seems to be rapidly rotated out and into the duplex without the formation of a specific complex. These studies could unravel a potentially comprehensive process of the repair protein to find and recognise the lesion through the slow kinetic pathway in which the more deformable damaged DNA is initially located by the protein; the protein subsequently compresses the duplex into an appropriate angle and direction to form a specific protein-DNA complex prior to being flipped and repaired.

Acknowledgements

I would like to take the opportunity to express my sincere gratitude to Drs. Charles A. Laughton and Stephen W. Doughty for giving me the opportunity to work on this project, and for all their valuable advice, continual guidance, kindness and inspiration throughout my course at the University.

I owe many thanks also to the numerous members of the molecular recognition group–Mark, Dan, Michele, Verity, Hao, Angelo, and Mike–for making such a pleasurable experience during my PhD course. I am truly indebted to Dr. Ian M. Withers, our postdoctoral fellow and the system administrator, for helping with the UNIX operating system, shell scripting and programming.

Finally, I would like to thank all my family and friends for their long-term encouragement. Particular gratefulness is going to my beloved wife for dedicating most of her time to support me and look after our two gorgeous sons. This work was generously supported by the Royal Thai Government Scholarship.

Contents

List of Figures

List of Tables

Abbreviations

5OHC	5-Hydroxycytosine
8OA	8-Oxoadenine
8OG	8-Oxoguanine
AMBER	Assisted Model Building with Energy Refinement
AP	Apurinic/apyrimidinic
APBS	Adaptive Poisson-Boltzmann Solver
BER	Base excision repair
CHARMM	Chemistry at Harvard Macromolecular Mechanics
COM	Centre-of-mass
DHT	Dihydrothymine
DHU	Dihydrouracil
DNA	Deoxyribonucleic acid
DSB	Double-strand break
dsDNA	Double stranded DNA
GAFF	General AMBER force field
GBSA	Generalised born surface area
FapydA	4,6-Diamino-5-formamidopyrimidine
FapydG	2,6-Diamino-4-hydroxy-5-formamidopyrimidine
FEP	Free-energy profile
ff	Force field
Fpg	Formamidopyrimidine-DNA glycosylase
H2TH	Helix-two turn-helix

HhH	Helix-hairpin-helix
hOGG1	Human oxoguanine DNA glycosylase 1
HR	Homologous recombination
kcal	Kilocalories
MD	Molecular dynamics
MM	Molecular mechanics
NER	Nucleotide excision repair
NHEJ	Non-homologous end joining
NMR	Nuclear magnetic resonance
PBC	Periodic boundary conditions
PCA	Principal component analysis
PDB	Protein Data Bank
Pol	Polymerase
QM	Quantum mechanics
PME	Particle Mesh Ewald
PMF	Potential of mean force
RESP	Restrained electronic potential
RC	Reaction coordinate
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
Sander	Simulated annealing with NMR-derived energy restraints
Tg	Thymine glycol
UDG	Uracil-DNA glycosylase
vdW	Van der Waal
VMD	Visual Molecular Dynamics
WC	Watson-Crick
WHAM	Weighted Histogram Analysis Method

Chapter 1

Introduction

1.1 DNA damage and cellular responses

Since Avery and co-workers reported that DNA is the material of inheritance for the first time in 1944 [2], DNA has been well acknowledged to be a biomolecule that encodes genetic information. The information is carried through genetic sequences of the four distinct bases adenine (A), cytosine (C), guanine (G) and thymine (T) along a DNA strand, which then are translated to the gene product (protein). Each base on one strand forms hydrogen bonds with its complementary base in standard Watson-Crick (WC) geometry, with A bonding only to T and G pairing only to C (Figure 1.1).

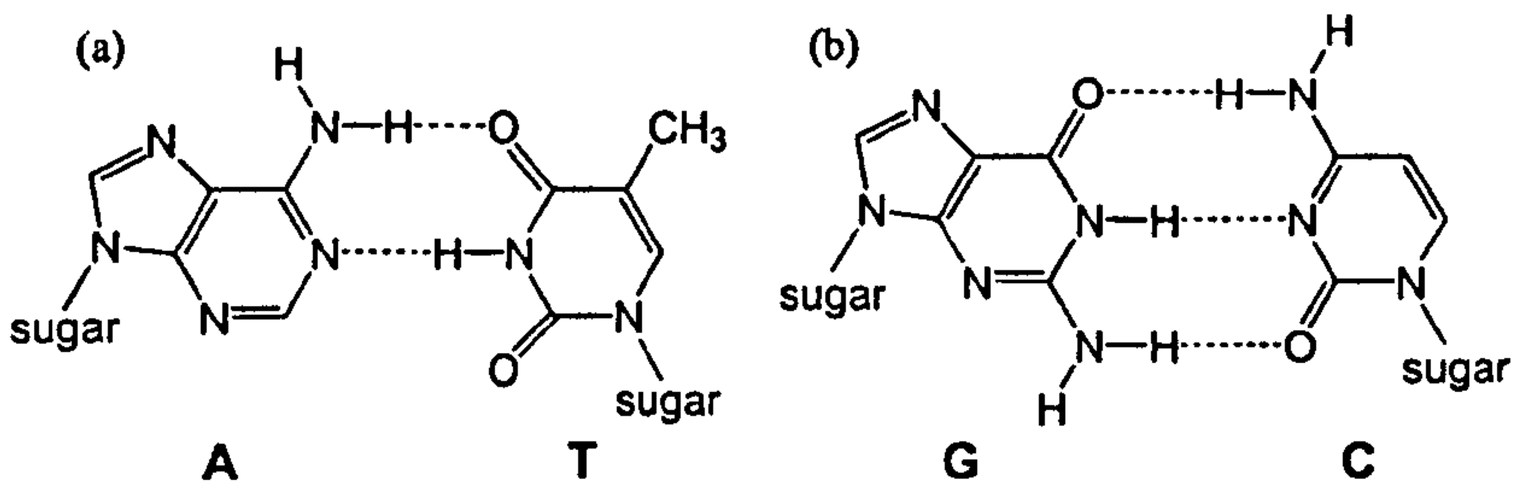


Figure 1.1: (a) A:T and (b) G:C base pairs in Watson-Crick hydrogen bonding pattern.

DNA suffers continual damage, due to endogenous and exogenous sources, either at the bases or in the phosphodiester backbone that the bases attach to.

In this study, DNA base damage will be predominantly considered rather than at other parts of DNA. DNA damage probably occurs at an incredible rate of up to 1 million lesions per cell per day [3] and can happen at any DNA sequence and in any stage of the cell cycle. If a damaged base is in a critical gene that controls cell proliferation, it can lead directly to cell death due to arrested DNA replication. DNA damage which occurs in a regulatory region of DNA such as a promoter or in a tumour suppressor gene results in a faulty gene or protein expression or increases the probability of the cancer cell formation, respectively. For example, when a DNA lesion is in a coding strand of DNA, a polymerase enzyme may stall at the damage leading to a truncated mRNA or may mistranscribe giving a wrong code in the mRNA. Due to these potentially deleterious effects, cells have evolved ingenious mechanisms for tolerating and repairing the damage. Failure in repair mechanisms can lead to hereditary diseases such as xeroderma pigmentosa (XP), hereditary non-polyposis colon cancer (HNPCC) and some forms of breast cancer [4].

DNA damage tolerance is to retain the progression of the cell cycle when DNA damage is present, thus allowing normal DNA replication and gene expression to bypass the unrepaired damage. Such bypassing sites of base damage results in the damage remaining in the cell. To proceed to normal cell division through the damage, specialised low-fidelity DNA polymerases are required for DNA replication whereas the high-fidelity polymerases typically stall at the damage. Low-fidelity polymerases are prone to generate replication mistakes or base mismatches, for instance, polymerase kappa (Pol κ) tends to insert adenine opposite an 8-oxoguanine lesion instead of cytosine during replication. Such base misincorporation leads to point mutations namely transitions and transversions. A transition mutation is when a pyrimidine is changed to another pyrimidine, or a purine to another purine, while when a purine is changed to a pyrimidine or *vice versa* this is called a transversion mutation. An example of a G:C $\rightarrow$ A:T transition after DNA replication is shown in figure 1.2.

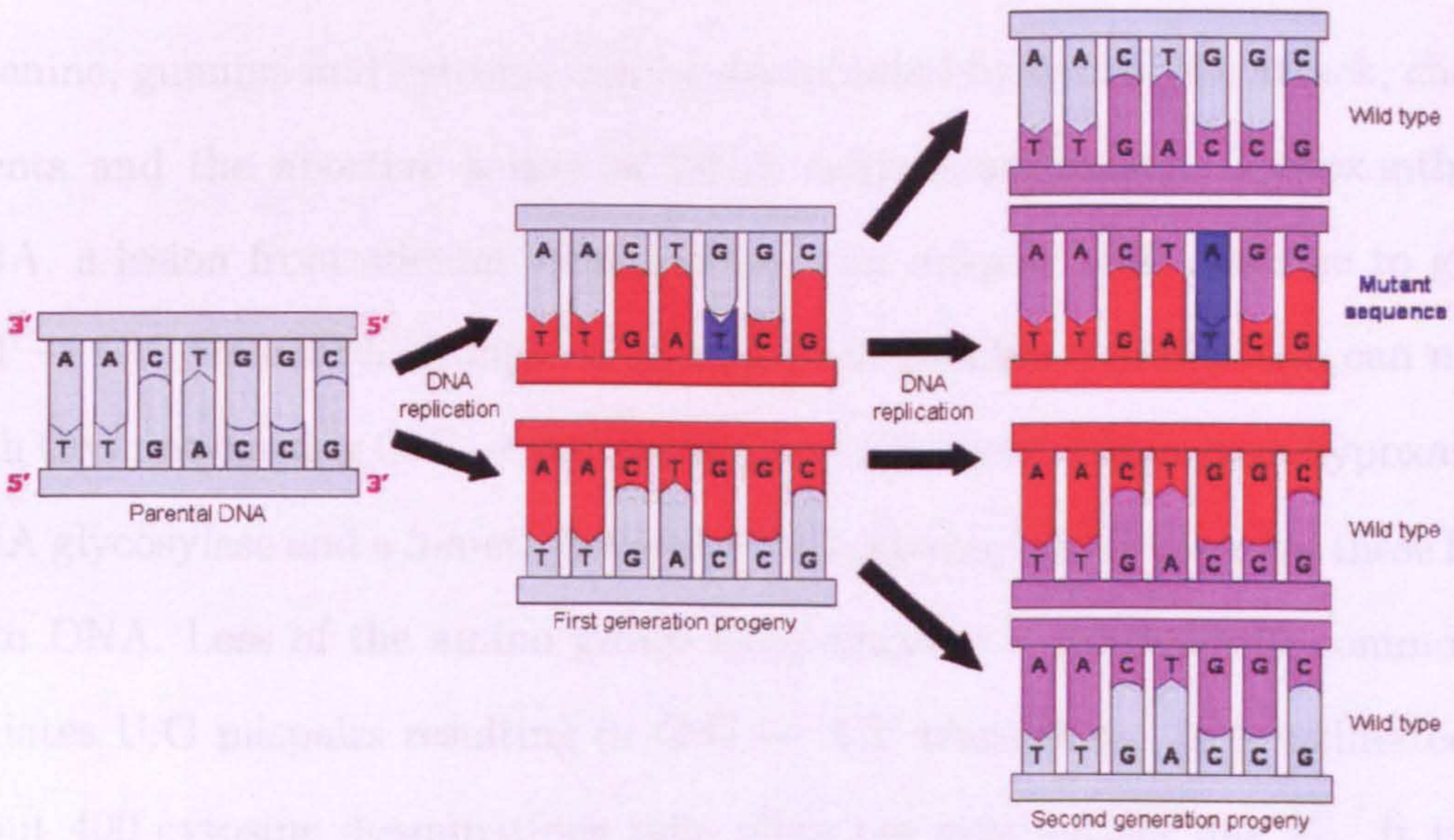


Figure 1.2: An example of a point mutation leading to G:C → A:T transition. The picture is reproduced from [5].

1.1.1 Types of DNA damage

Exposure of DNA to various sorts of mutagens produces multiple forms of damage through several different types of processes. There are five types of DNA damage consisting of hydrolysis, deamination, alkylation, dimerisation and, of particular interest in this study, oxidation (Figure 1.3).

1.1.1.1 Hydrolysis

Spontaneous hydrolysis cleaves the chemical bond between a DNA base and its respective deoxyribose leading to abasic sites known as AP sites (apurinic/apyrimidinic). In mammalian cells, it is estimated that the loss of purine bases (guanine and adenine) is at the rate of about 10,000 bases per cell cycle, while the rate of pyrimidine bases (cytosine and thymine) lost is noticeably slower, resulting about 500 bases lost per cell cycle [6]. These AP sites are nevertheless a general intermediate in the base excision repair (BER) pathway as they are subsequently recognised and cleaved by AP endonucleases. The efficient repair of these AP sites is crucial because unrepaired AP sites block the progress of DNA replication and cause mutagenesis and eventually cell death [7, 8]

1.1.1.2 Deamination

Adenine, guanine and cytosine can be deaminated by hydrolytic attack, chemical agents and the abortive action of DNA methyltransferases. Hypoxanthine in DNA, a lesion from adenine deamination, can mispair with cytosine to give an A:T → G:C transition. Xanthine, arising from guanine deamination, can mispair with thymine causing G:C → A:T transition mutations. There are a hypoxanthine DNA glycosylase and a 3-methyladenine DNA glycosylase II to excise these lesions from DNA. Loss of the amino group from cytosine is particularly common and initiates U:G mispairs resulting in G:C → A:T transitions. It is estimated that about 400 cytosine deaminations take place per genome per day [6]. It is clear that these lesions have to be repaired by specific DNA repair enzymes prior to the next round of DNA replication to prevent mutagenesis.

1.1.1.3 Alkylation

DNA alkylation occurs when nucleobases are attacked by alkylating agents such as cytotoxic chemotherapy agents e.g. nitrogen mustard compounds and endogenous agents such as *S*-adenosylmethionine, a biological methyl donor. Alkylated bases, such as O^6-methylguanine, N^1-methyladenine and N^3-methylcytosine, have an altered hydrogen bonding pattern which can lead to base mispairing. Of these alkylated bases, O^6-methylguanine, which can mispair with thymine, and is strongly mutagenic by inducing G:C → A:T transition mutations. The lesion is directly repaired in an irreversible reaction by a suicide repair enzyme O^6-alkylguanine alkyltransferase (AGT) or O^6-methylguanine-DNA methyltransferase (MGMT) [9].

1.1.1.4 Dimerisation

Two adjacent pyrimidine bases are able to be crosslinked due to ultraviolet (UV) radiation. Cyclobutyl pyrimidine dimers (CPDs) and pyrimidine (6-4) pyrimidinone dimers are two major UV-induced DNA lesions causing skin photocar-

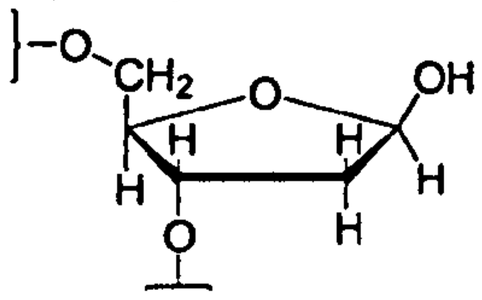
cinogenesis [10]. These lesions can be repaired by DNA photolyases, base and nucleotide excision repair enzymes.

1.1.1.5 Oxidation

Oxidative DNA damage is generated when DNA bases are exposed to reactive oxygen species (ROS), which include the highly reactive hydroxyl radical ($^{\bullet}OH$), superoxide radical ($O_2^{\bullet -}$) and non-radical hydrogen peroxide (H_2O_2). Cell metabolism, UV radiation and chemicals can all increase the level of ROS in cells leading to the creation of several oxidation products such as 5-hydroxycytosine, thymine glycol, 8-oxopurine and formamidopyrimidine lesions. Of these lesions, oxidative guanine damage is the most extensively studied because it is easily measured [11] and guanine is the most easily oxidised among the DNA bases because of its lowest oxidation potential [12, 13]. It has been estimated that about 7500 8-oxo-7,8-dihydroguanine (8OG in keto form or known as 8-hydroxyguanine in enol form) lesions are yielded from hydroxyl radical attack in each human cell [14]. Oxidative bases are primarily repaired by base excision repair, and also by nucleotide excision repair to a lesser extent. Unless repaired, oxidative DNA damage can result in mutagenesis, carcinogenesis and aging [15].

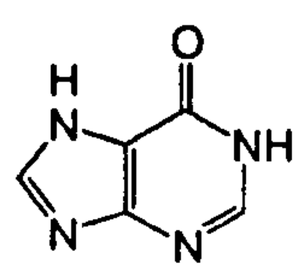
Oxidative damage to DNA contributes to the incidence of several serious diseases such as cancer, chronic inflammatory diseases, cardiovascular diseases, aging and neurodegenerative problems [16]. Numerous studies have attempted to explore a relationship between levels of oxidative lesions and occurrences of particular diseases (see review in [17]). The first report on an association of oxidised bases with breast cancer by Malins and Haimanot in 1991 demonstrated a 9-fold elevation in the levels of 8-oxopurine and formamidopyrimidine lesions in the tumour tissue compared with neighbouring normal tissue [18]. Subsequently, Shimoda *et al.* reported that the levels of 8OG in liver tissue with chronic hepatitis are increased compared with control liver [19]. Thus, it is highly likely that the level of oxidative DNA damage is important in the formation of various diseases.

a) Hydrolysis

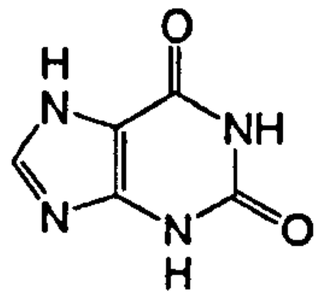


AP site

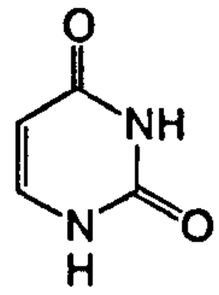
b) Deamination



hypoxanthine

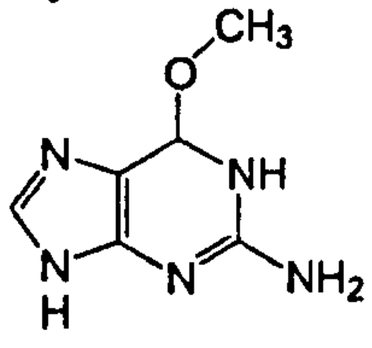


xanthine

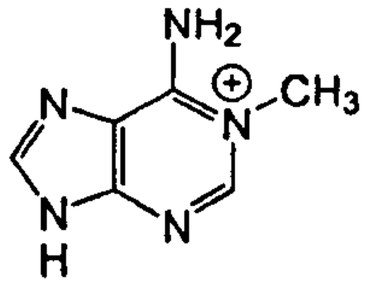


uracil

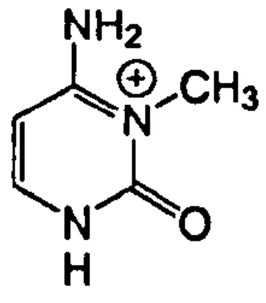
c) Alkylation



O^6-methylguanine

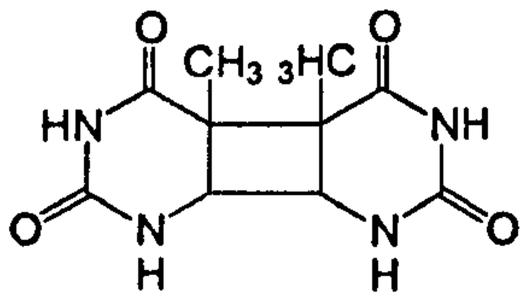


N^1-methylguanine

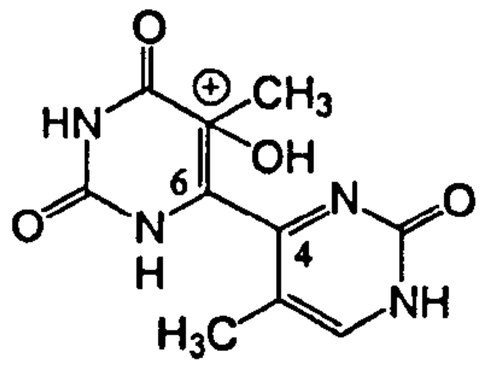


N^3-methylcytosine

d) Dimerisation

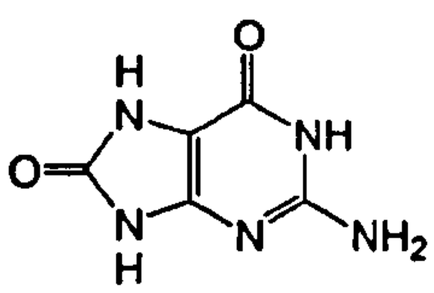


thymine dimer

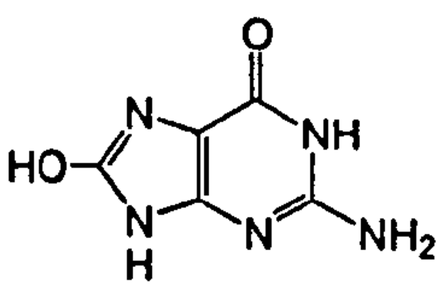


thymine-thymine pyrimidine (6-4) pyrimidone dimer

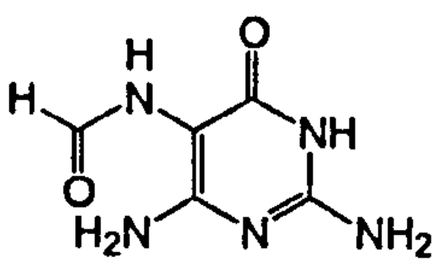
e) Oxidation



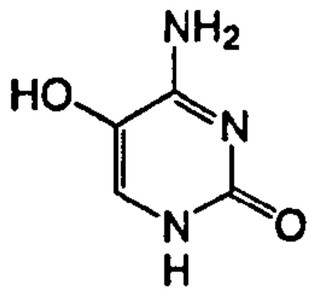
8-oxo-7,8-dihydroguanine



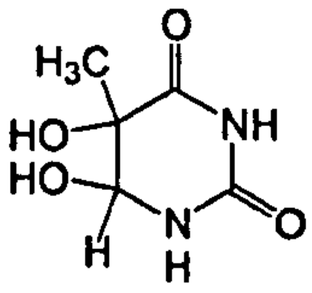
8-hydroxyguanine



2,6-diamino-4-hydroxy-5-formamidopyrimidine



5-hydroxycytosine



Thymine glycol

Figure 1.3: Chemical structures of major DNA damage classified by chemical modifications.

1.1.2 Cellular responses to DNA damage

When the DNA structure was first reported in 1953 by James Watson and Francis Crick, it was supposed to be an extremely stable double helix. Twenty one years later, however, Crick published his opinion on the possibility of the existence of a DNA repair system to retain genetic information [20]. Nowadays it is unsurprising that cells have evolved multiple biochemical processes to cope with the vast diversity of DNA damage. Responses to DNA damage depend on the damage type. On DNA base damage, it alters not only the chemical structure of nucleobases but also the spatial conformation of DNA helix that cells should be able to detect and repair such modifications specifically. DNA repair processes mainly involve one of two fundamental mechanisms: excision of damaged elements, and reversal of DNA damage [21]. In addition to the base damage, single- and double-strand breaks are another type of DNA damage. Although the strand breaks do not directly change the genetic information, if not repaired they can interfere in DNA replication. Moreover, in cases of unrepaired or unrepairable damage, cells have developed a process known as translesion DNA synthesis to bypass the lesions by particular enzymes.

1.1.2.1 Excision of base damage

If there is only one damaged strand in the two strands of a double helix, the undamaged strand can be used as a template to direct the correction of the damaged strand. To remove the damaged base from the paired bases of DNA, it can be excised from the duplex as a free base or as a nucleotide, and be replaced with an undamaged nucleotide complementary to that found in the undamaged strand. These repair processes are biochemically and mechanistically distinct and are referred to base excision repair and nucleotide excision repair, respectively. The last excision mechanism is called mismatch repair, particularly for the removal of mispaired bases in DNA.

Base excision repair (BER) is a primary repair mechanism to restore genetic information [22]. The main sources of lesions repaired by BER pathway are from oxidation, non-enzymatic alkylation, hydrolysis, deamination, and including incorporated uracil. The mechanism of BER is shown in figure 1.4. The first step in BER is conducted by recognition of the lesions by specific enzymes, the DNA glycosylases. DNA glycosylases are normally responsible for recognition and removal of the damaged bases by flipping the lesion out from the duplex and hydrolysing the *N*-glycosidic bond leaving an AP site. The phosphodiester bond of an AP site is subsequently cleaved by an AP endonuclease (APE) giving rise to a single-strand break whereas some DNA glycosylases also possess AP endonuclease activity. Formamidopyrimidine DNA glycosylase (Fpg or MutM), for example, is a bifunctional repair enzyme which is responsible for removing a FapydG (2,6-Diamino-4-hydroxy-5-formamidopyrimidine) lesion. The correct nucleotide is then inserted by DNA polymerase β (Pol β). Further repair is to replace one (short-patch pathway) or more (long-patch pathway) nucleotides, depending on the structure of 5'-deoxyribose-5-phosphate (5'dRP) after being excised by APE or DNA glycosylases. In the short-patch pathway, 5'dRP, which retains the hemiacetal sugar form, is then cleaved by Pol β and the DNA gap is then ligated by a X-ray cross-complementing 1 (XRCC1)-DNA ligase III complex. Alternatively, the 5'dRP ring is cleaved by endonuclease III through a β-elimination mechanism leaving a 3'-α,β-unsaturated aldehyde sidechain [23]. Long-patch repair is responsible for dealing with the aldehyde form of 5'dRP and is a proliferating cellular nuclear antigen (PCNA)-dependent pathway which requires multiple enzymes such as replication factor C (RF-C), flap endonuclease (FEN1), Pol δ or Pol ϵ and DNA ligase I to complete the repair process. It is thought that most of the lesions which are recognised by DNA glycosylases are repaired via the single-nucleotide replacement pathway, whereas AP sites from the spontaneous hydrolysis of bases are repaired by long-patch BER [24].

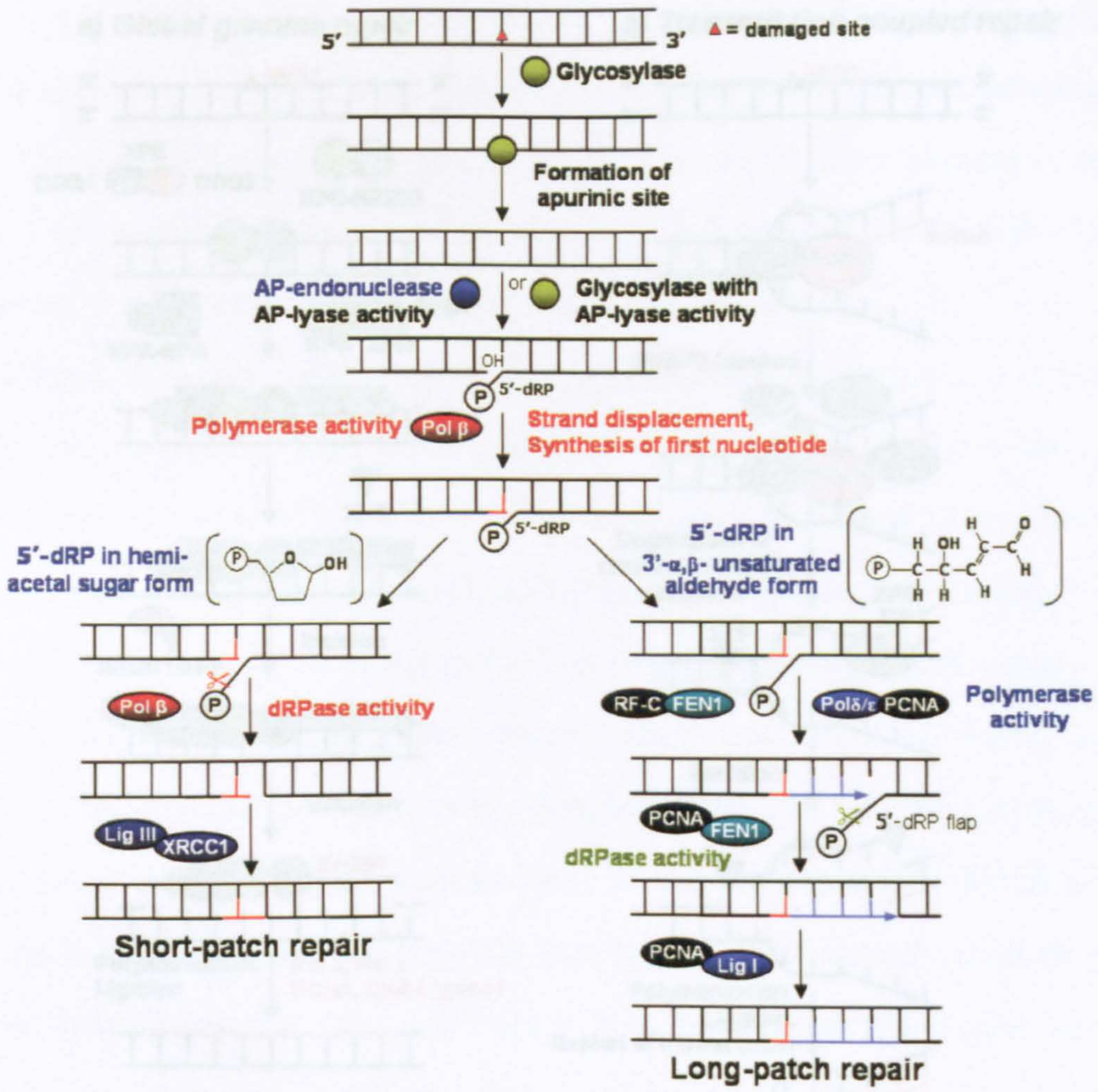


Figure 1.4: Mechanism of base excision repair (BER). The picture is reproduced from [25].

Nucleotide excision repair (NER) is responsible for repairing bulky, helix-distorting changes such as thymine dimers as well as single-strand breaks. The NER process is moreover believed to be a supportive system for BER to remove oxidative DNA damage [26]. There are two NER sub-pathways based on the damage recognition step: global genome repair (GGR) and transcription-coupled repair (TCR) as shown in figure 1.5. GGR is initiated by lesion recognition by the xeroderma-pigmentosum complementation group C (XPC)-Rad23B, replication protein A (RPA)-XPA or damaged DNA binding protein (DDB) complexes [25]. Each complex has a different binding affinity to each type of damage, for example, the DDB complex involves the recognition of CPDs with high efficiency, while the

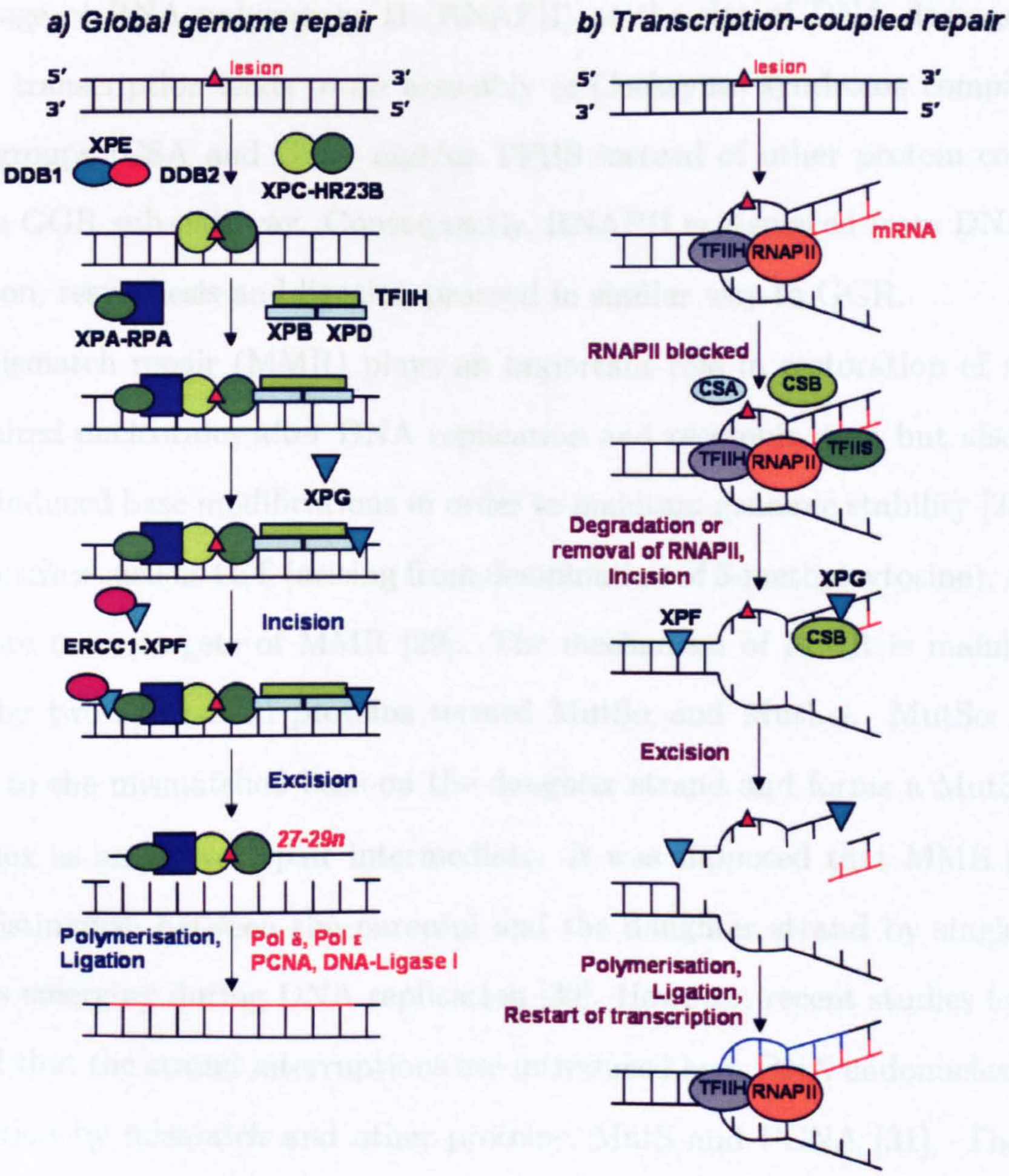


Figure 1.5: Mechanism of nucleotide excision repair (NER). The picture is reproduced from [25].

XPC-Rad23B specifically recognises pyrimidine (6-4) pyrimidinone dimers. After damage is recognised, the transcription factor TFIIH, which possesses DNA helicase activity, is recruited to the site of DNA damage to unwind the DNA around the lesion [27]. Subsequent dual incisions of 3'-flanked and 5'-flanked to the lesion are performed by XPG and the excision repair cross-complementing group 1 (ERCC1)-XPF complex, respectively, resulting in an oligonucleotide fragment of 27-29 nucleotides. Hence the damage is removed as part of the fragment. Finally, the gap is filled in by DNA synthesis with Pol δ and ϵ using the undamaged

strand as a template and sealed by DNA ligase I. In transcription-coupled NER, blockage of RNA polymerase II (RNAPII) at the site of DNA damage during DNA transcription leads to an assembly of Cockaynes syndrome complementation groups (CSA and CSB) and/or TFIIS instead of other protein complexes in the GGR sub-pathway. Consequently, RNAPII is displaced from DNA. DNA excision, resynthesis and ligation proceed in similar way to GGR.

Mismatch repair (MMR) plays an important role in restoration of not only mispaired nucleotides after DNA replication and recombination but also chemically induced base modifications in order to maintain genomic stability [28]. Base mismatches such as G:T (arising from deamination of 5-methylcytosine), A:C and C:C are main targets of MMR [29]. The mechanism of MMR is mainly mediated by two specialised proteins termed MutSα and MutLα. MutSα initially binds to the mismatched base on the daughter strand and forms a MutSα-ADP complex as an active repair intermediate. It was supposed that MMR proteins can distinguish between the parental and the daughter strand by single-strand breaks emerging during DNA replication [30]. However, recent studies have suggested that the strand interruptions are introduced by a DNA endonuclease upon activation by mismatch and other proteins, MutS and PCNA [31]. The active complex, which has ATPase activity, enables translocation of the complex along DNA and association with MutLα. The single-strand breaks on the daughter DNA are used as starting points for exonuclease I to remove the mispaired base. DNA resynthesis and religation are performed by Pol δ and DNA ligase.

1.1.2.2 Direct damage reversal

Some DNA lesions are repaired by directly reversing them to the original bases. Such direct reversal mechanisms are specific to only three types of base damage: thymine dimers, O^6-methylguanine, and certain methylations of adenine and cytosine. The enzyme photolyase can instantly reverse thymine dimers to restore two adjacent thymine bases. O^6-methylguanine is irreversibly repaired by AGT

in a stoichiometric mechanism. AGT must flip the target nucleotide into the binding pocket and transfer the methyl group to an active residue cysteine [32], leaving a normal guanine in DNA. The last type of DNA damage reversed by this direct mechanism is N^1-methyladenine and N^3-methylcytosine. These cytotoxic lesions are repaired by AlkB enzymes via an oxidative mechanism resulting in the reversion to the normal base and the release of the methyl group as formaldehyde [33, 34].

1.1.2.3 Double-strand breaks repair

When DNA is attacked by free radicals or ionising radiation the result is frequently double-strand breaks (DSBs) that need repairing. DSBs additionally occur during the DNA replication process. Although DSBs do not directly alter genetic information, the fracture of the genome can interfere with normal DNA transactions and lead to genome rearrangements that certainly endanger cells. To repair DSBs, there are two principal mechanisms termed non-homologous end joining (NHEJ) and homologous recombination (HR). NHEJ is a relatively simple process by joining two broken DNA ends together without the requirement of a homologous template (see figure 1.6). The repair is initiated by binding of the Ku70 and Ku80 heterodimer as a ring surrounding one broken end. DNA-dependent protein kinase catalytic subunit (DNA-PKcs) subsequently binds to the heterodimer in order to bring the two ends in close proximity and catalyses the joining of the two DNA ends [35]. The NHEJ process is eventually completed by XRCC4-DNA ligase IV complex which directly joins the two broken ends [36].

Alternatively, recombinational repair is another mechanism to repair DSBs, in which the presence of an identical or nearly identical sequence is required to be used as a template. HR is a slow repair process but it is important in the repair of DSBs that occur during DNA replication [35]. As shown in figure 1.7, HR repair begins with nucleolytic resection of the DSBs by the Mre11/Rad50/Nbs1 protein complex resulting a 3' single-stranded DNA fragment. A heptameric

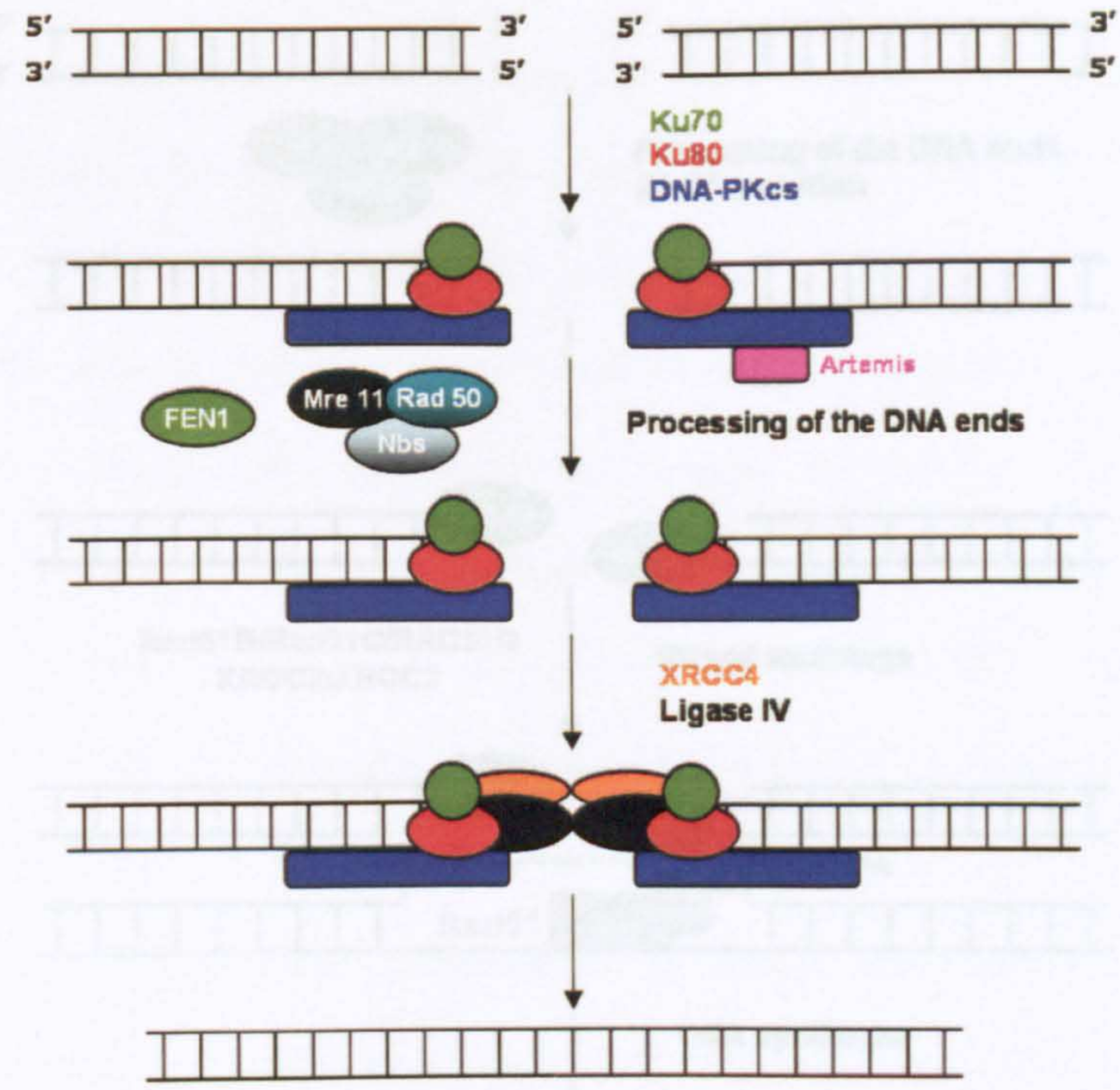


Figure 1.6: Mechanism of non-homologous end joining (NHEJ). The picture is reproduced from [25].

ring complex formed by Rad52 proteins binds to the 3' overhang to protect the ssDNA against exonucleolytic digestion [25]. Rad52 therefore interacts with a Rad51 nucleoprotein filament causing an exchange of the broken DNA end with the homologous DNA template. The crossing over DNA strand is completed after DNA synthesis, ligation and branch migration as in the classical model of Holliday junction [37].

1.1.2.4 Translesion DNA synthesis

Translesion DNA synthesis (TLS) is an ability of particular DNA polymerases to replicate through DNA damage rather than terminate the cell cycle when encounter the lesion. This process involves the use one of a group of TLS polymerases, most of which belong to the Y-family [38], to insert nucleotides opposite DNA lesions. Unfortunately, TLS polymerases have considerably low fidelity for nucleotide insertion that can bypass the damage in error-free or error-prone

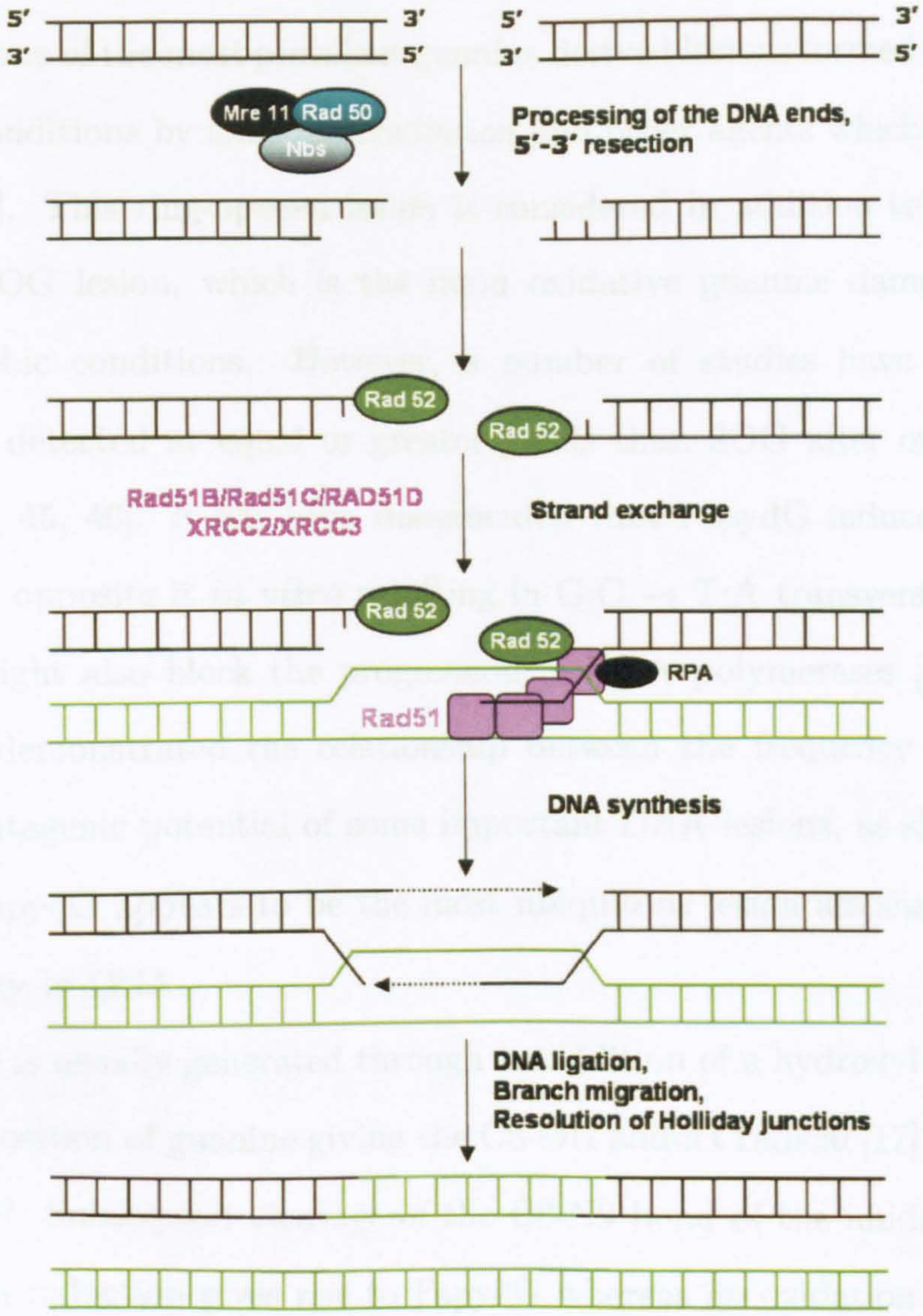


Figure 1.7: Mechanism of homologous recombination (HR). The picture is reproduced from [25].

mechanism, which the latter promoting mutagenesis. For example, Pol η is able to insert adenine opposite cyclobutane thymine dimers with similar efficiency to undamaged DNA [39], whereas Pol ζ inserts guanine opposite the lesion site producing mutations.

1.2 Molecular recognition of FapydG by Fpg

FapydG is one of the most prevalent guanine derived lesions formed under oxygen-deficient conditions by ionising irradiation and other agents which produce ROS [40, 41, 42]. This ring-opened lesion is considered in addition to the well documented 8OG lesion, which is the main oxidative guanine damage developed under aerobic conditions. However, a number of studies have revealed that FapydG is detected at equal or greater yields than 8OG after oxidative stress [42, 43, 44, 45, 46]. It has been documented that FapydG induces misincorporation of A opposite it *in vitro* resulting in G:C → T:A transversion mutations [47] and might also block the progression of DNA polymerases [48]. A recent study has demonstrated the relationship between the frequency of occurrence and the mutagenic potential of some important DNA lesions, as shown in figure 1.8 [22]. FapydG appears to be the most ubiquitous lesion associated with high mutagenicity in DNA.

FapydG is usually generated through an addition of a hydroxyl radical (•OH) at the C8 position of guanine giving the C8-OH adduct radical [17], as illustrated in figure 1.9. Subsequent cleavage of the C8-N9 bond of the imidazole ring and one-electron reduction gives rise to FapydG whereas an oxidation of the adduct yields 8OG. The C8-OH adduct radical may also undergo the reduction prior to ring cleavage leading to the formation of 7,8-dihydro-8-hydroxyguanine, which is then transformed to yield FapydG.

Although the formation of both FapydG and 8OG lesions is through the same intermediate as shown in figure 1.9, the remarkable changes in the structure of FapydG from the original guanine base, where the cleavage of the aromatic imidazole resulting in the increased conformational flexibility of *N*-glycosidic bond and formamide group, are fascinating questions to understand the effects on DNA structure and how the lesion is recognised by DNA repair enzymes. Unfortunately, the first problem in studying FapydG is the difficulty to produce pure oligonucleotides containing a β-FapydG lesion, which is supposed to be a natu-

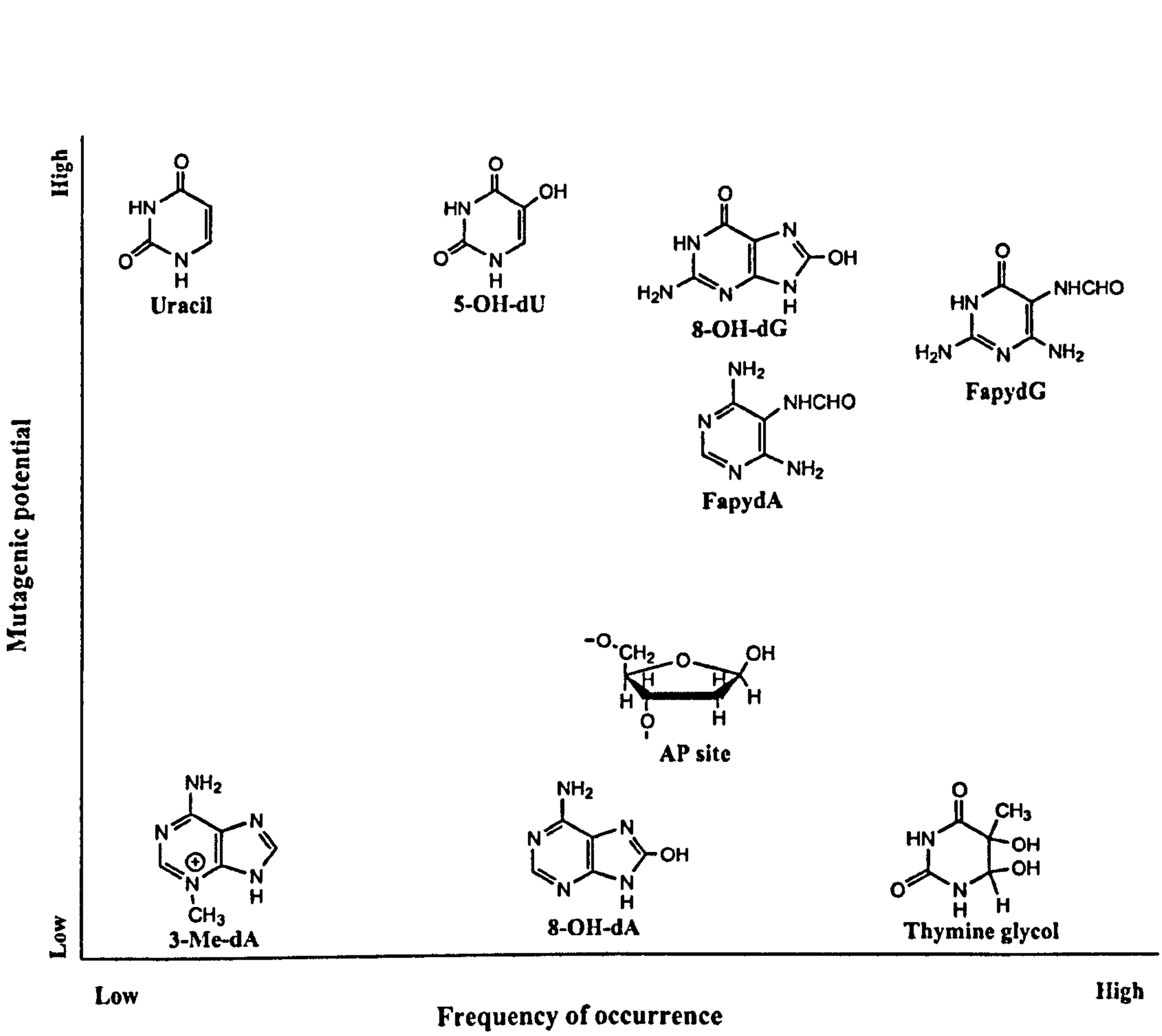


Figure 1.8: Frequency of occurrence of some important forms of DNA damage and their mutagenic potential. Chemical structures of the damage with their name are shown. The figure is reproduced from [22].

rally occurring anomer in the DNA duplex [49]. Chemically synthetic β-FapydG tends to rapidly anomerise to give α-FapydG under conditions required for DNA synthesis [50]. Thus, there has been relatively little information on FapydG. Two types of FapydG analogues, however, have been used as a substitute for FapydG to be inserted site-specifically into DNA as the pure β-anomer, as shown in figure 1.10.

The first analogue is a nonhydrolysable FapydG with a *C*-glycosidic bond (β-C-FapydG) [51]. This β-C-nucleoside analogue is supposed to be an inhibitor for Fpg since Fpg can tightly bind to DNA containing β-C-FapydG pairing to C *in vitro* but is unable to excise the damage [52]. Another analogue is a carbocyclic FapydG (*c*FapydG), which contains a cyclopentane ring instead of the 2'-deoxyribose sugar [53]. The crystal structure of a bacterial Fpg enzyme, *Ll*Fpg from *Lactococcus lactis*, bound to a *c*FapydG-containing duplex was solved by

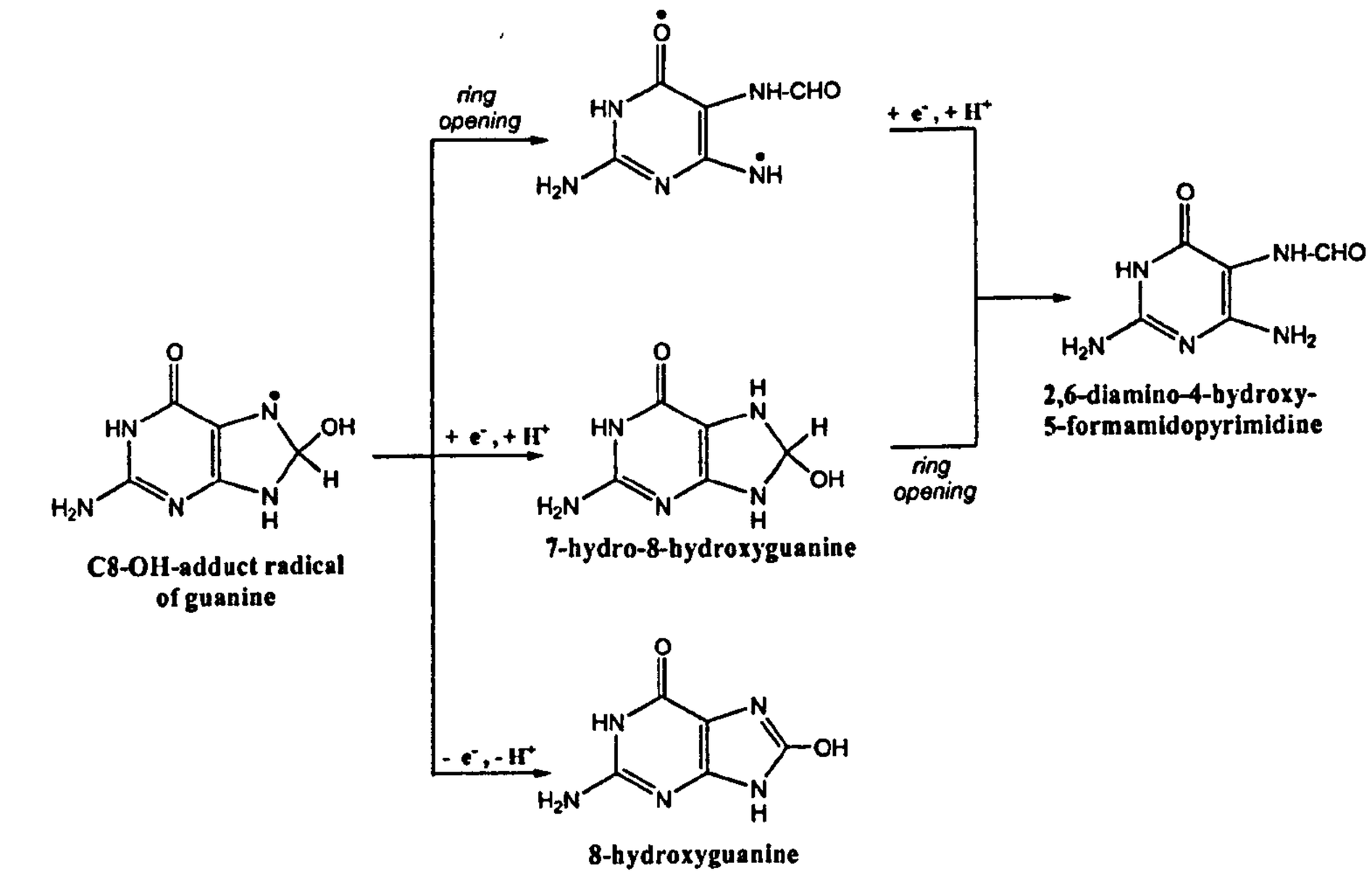


Figure 1.9: Formation of FapydG and 8OG through the C8-OH adduct radical of guanine in the absence of oxygen. The figure is reproduced from [17].

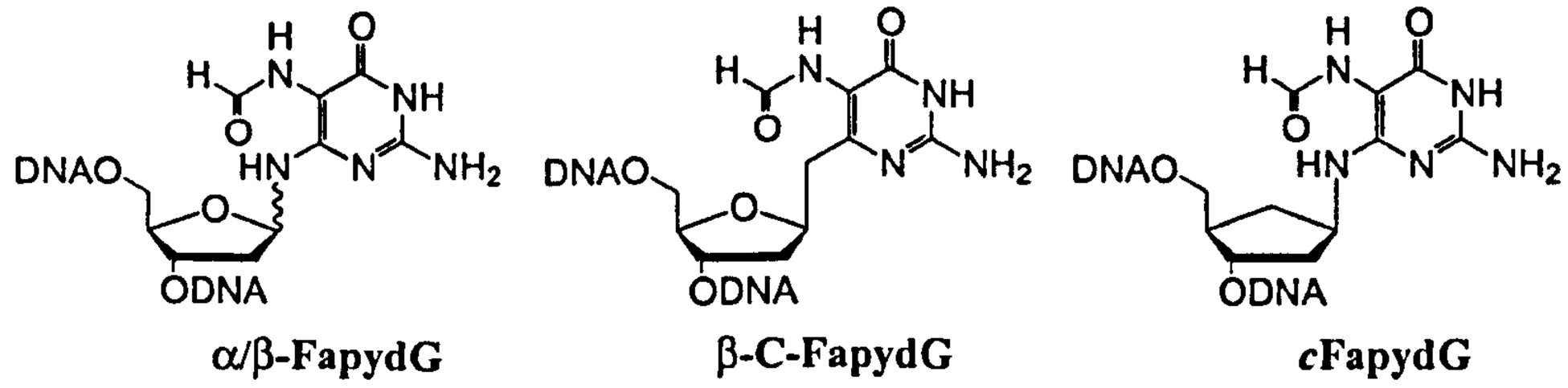


Figure 1.10: Chemical structures of FapydG analogues.

Coste *et al.* in 2004 [54]. This noteworthy study provides us with a structural basis for recognition of FapydG by Fpg which is a prime step in the BER process. The structure was deposited in the Protein Data Bank (PDB) [55] with two PDB entries, 1TDZ (solved at 1.8 Å resolution with a missing loop on the residues 220 to 224) and the complete model with 1.95 Å resolution (PDB code 1XC8, released one year after 1TDZ).

The global structure of the *Ll*Fpg-DNA complex (PDB code 1XC8) is illustrated in figure 1.11. *Ll*Fpg consists of two globular domains, N- and C-terminal domains, with a pronounced cleft between the domains [56]. The N-domain

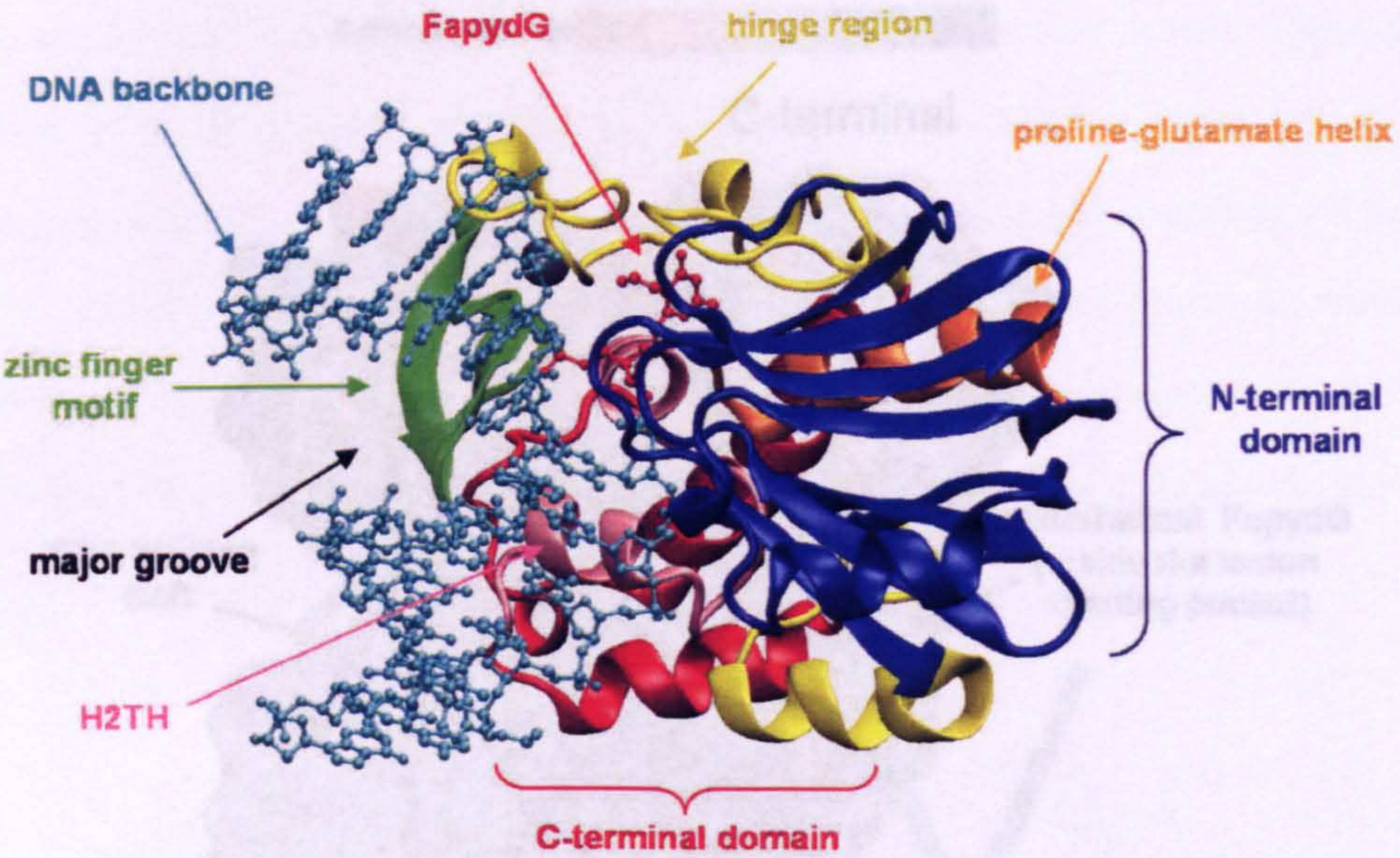


Figure 1.11: An overview of *c*FapydG-containing DNA bound to Fpg. Overall structure of Fpg showing a proline-glutamate helix coloured in orange, the N-terminal domain in dark blue, the hinge region in yellow, the C-terminal domain in red and pink, the zinc finger subdomain in green, the DNA backbone in a small cyan ball and stick model, and FapydG residue in red. The picture is modified from [54].

protein is composed of a proline-glutamate helix, followed by two 4-stranded β sheets that form an antiparallel β sandwich. The C-terminal is divided into two subdomains: an α-helix rich domain containing the helix-two turn-helix (H2TH) motif and a zinc finger domain with β-hairpin loop which intercalates into the DNA major groove. Though Fpg has a completely different molecular scaffold from other members of its DNA glycosylase superfamily that possess the helix-hairpin-helix (HhH) motif, Fpg functions the same as other glycosylases in the BER pathway. The cleft or groove formed between these two domains contains an abundance of positively charged residues providing an electrostatically positive surface for binding to the negatively charged of the DNA backbone. Figure 1.12 illustrates an electrostatic potential surface of the Fpg protein generated in Visual Molecular Dynamics software (VMD) by an Adaptive Poisson-Boltzmann Solver (APBS) approach [57]. The *c*FapydG-containing duplex fits nicely into the groove and exhibits a distorted structure with an extrahelical position of the

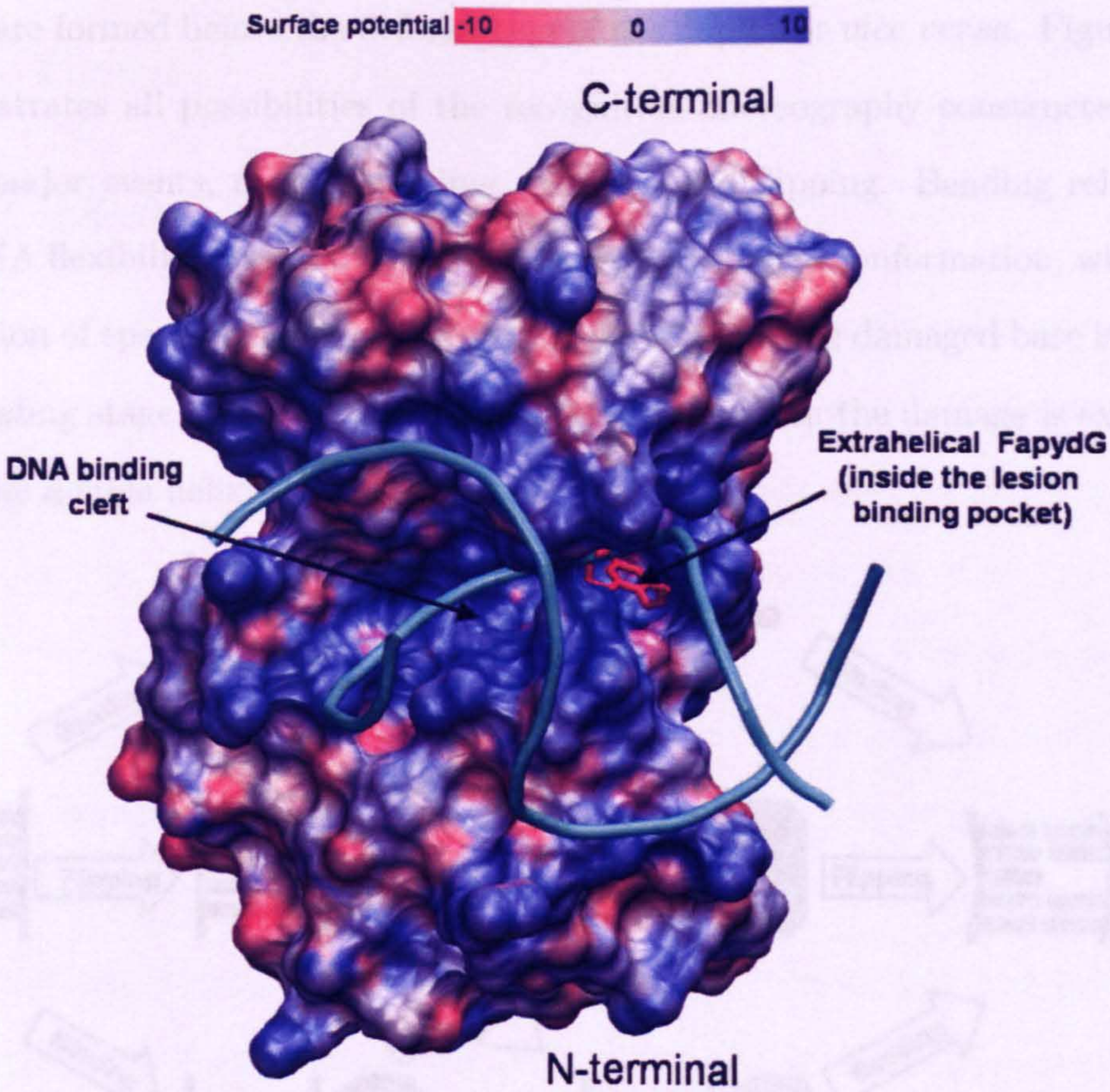


Figure 1.12: Structure of the Fpg-DNA complex. Solvent-accessible surface area of Fpg coloured according to electrostatic potential demonstrating the highly positive DNA-binding cleft of the enzyme. The DNA backbone is represented in a cyan rod, and the extrahelical FapydG residue in a red licorice model buried inside the lesion binding pocket.

damaged base buried inside the enzymatic binding pocket. The crystal structure has revealed that the DNA duplex is sharply kinked towards the major groove at the lesion site. To summarise from the *Ll*Fpg-DNA complex, there are three important features for DNA damage recognition: the protein specifically binding to the lesion, bending the DNA centred on the damaged base, and flipping the damage into the pocket, as being usually observed in all DNA glycosylase-DNA complexes studied to date [58].

Nonetheless, dynamic aspects of three stages of DNA damage recognition are doubtful. For example, it is not clear whether the flipped out stage of the

lesion are formed before the deformation of the duplex or *vice versa*. Figure 1.13 demonstrates all possibilities of the recognition choreography constructed from three major events, namely bending, binding, and flipping. Bending relates to the DNA flexibility required to obtain the sharply kinked conformation, while the formation of specific contacts between the protein and the damaged base is called the binding stage. Flipping is the process through which the damage is extruded from the double helix.

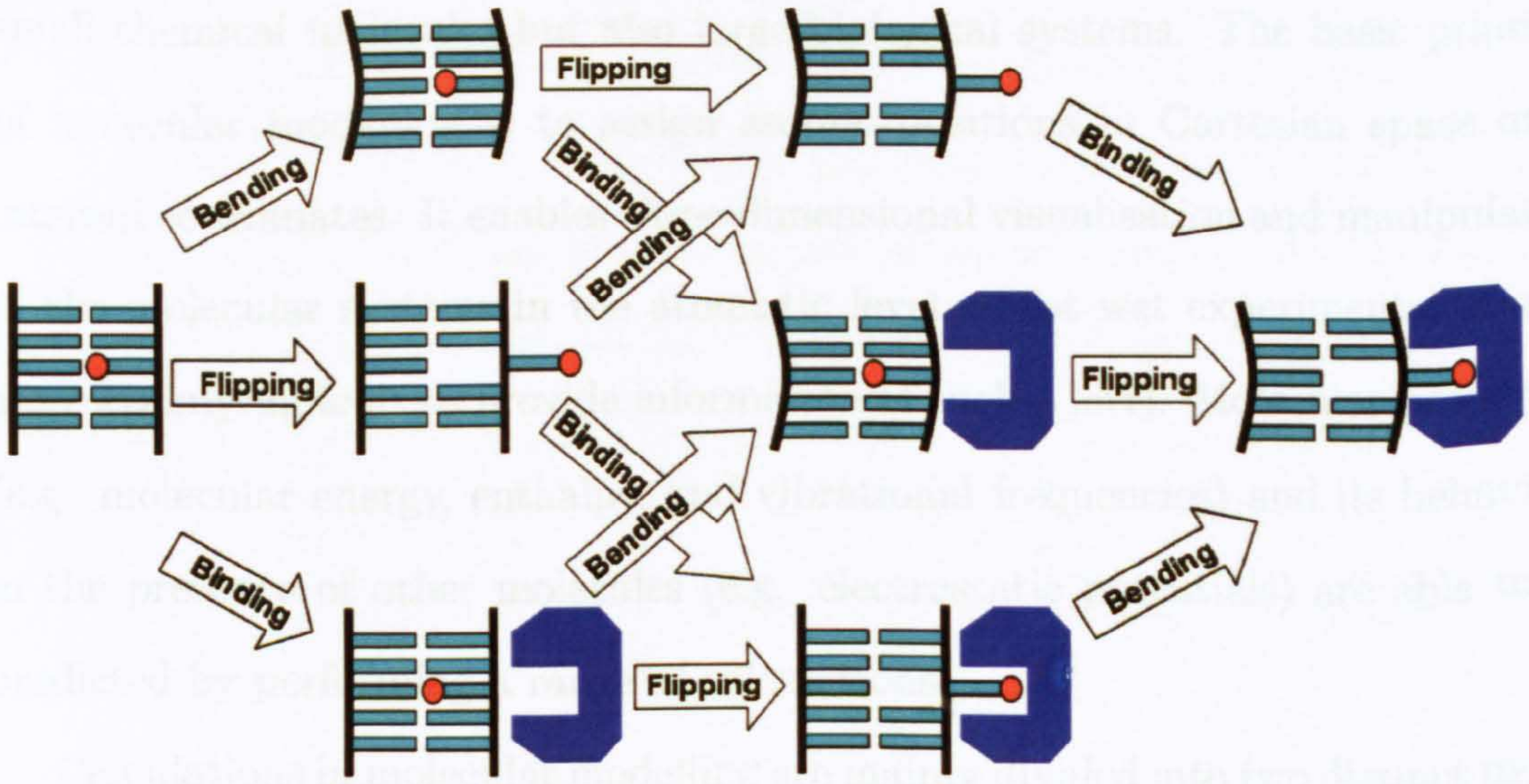


Figure 1.13: Recognition choreography. Possible recognition pathways from the damage occurred through the lesion extruded into the protein binding pocket. There are six possible recognition pathways: Bending → Flipping → Binding, Bending → Binding → Flipping, Flipping → Bending → Binding, Flipping → Binding → Bending, Binding → Bending → Flipping, Binding → Flipping → Bending.

Without an understanding of the dynamic behaviour of each event in the presence and absence of the lesion, it is certainly difficult to answer how DNA glycosylases can find and distinguish the lesion in a large excess of undamaged bases. One powerful approach to gain insights into dynamic aspects of those biomolecules are molecular modelling and dynamics simulations.

1.3 Molecular modelling and dynamics simulations

1.3.1 Molecular modelling

Molecular modelling is a generic term that refers to computational techniques which are used to depict, describe, or evaluate any aspect of the properties or structure of a molecule [59]. The technique is widely used for studying not only small chemical molecules but also large biological systems. The basic principle of molecular modelling is to assign atomic positions in Cartesian space or in internal coordinates. It enables three-dimensional visualisation and manipulation of the molecular systems in the atomistic level whilst wet experimental studies are certainly difficult to provide information of such a level. Molecular properties (e.g. molecular energy, enthalpy, and vibrational frequencies) and its behaviour in the presence of other molecules (e.g. electrostatic potentials) are able to be predicted by performing a range of calculations.

Calculations in molecular modelling are mainly divided into two distinct methods: quantum mechanical and molecular mechanical methods. Quantum mechanics (QM) is principally based on the Born-Oppenheimer approximation that nuclei and electrons are separated from each other. The atomic nuclei are considered as point charges where electrons are explicitly designated. The evaluation of nuclear-electron and electron-electron interactions are used to solve the spatial distribution of nuclei and electrons and their energies. One definite application of using QM is to estimate the reaction pathway of a chemical bonding process. Consequently, the QM method is broadly used to theoretically investigate structural and thermodynamic data, transition-state formation and force field parameterisation. Unfortunately, this sophisticated approach is extremely computationally time consuming and it is limited to systems of only a hundred or so atoms, and so is unable to deal with biomolecular systems such as protein-DNA complexes.

In contrast to QM, molecular mechanics (MM), also known as force field methods, typically ignores the electronic motions and treats an individual atom as a point charge with an associated mass. This mechanical approach significantly reduces computational time, making it possible to quantify molecular properties on systems containing thousands of atoms. Calculation of the analytical potential energy (E_{MM}), one of the most interesting molecular properties, is performed by adding up bonded energy terms that define the deviation of bond lengths, bond angles and torsion angles away from their equilibrium values and non-bonded energy terms that describe van der Waals and electrostatic interactions. All standard values of equilibrium bond lengths and bond angles, partial charges, force constants and van der Waals parameters are known as a force field. For biomolecular models, the assisted model building and energy refinement (AMBER) force field [60] is widely used to calculate E_{MM} as shown in equation 1.1.

$$E_{MM} = \sum_{\text{bonds}} K_r (r - r_{eq})^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2} (1 + \cos[n\phi - \gamma])$$

$$+ \sum_{i<j}^{\text{atoms}} \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \sum_{i<j}^{\text{atoms}} \frac{q_i q_j}{\epsilon R_{ij}} \tag{1.1}$$

where $r - r_{eq}$ and $\theta - \theta_{eq}$ are the deviation of a bond length or angle from a minimum energy value respectively, with K_r and K_θ representing the spring constant in accordance with Hooke's law. The approximation of the bond energies exhibits a simple harmonic energy curve where the energy quadratically increases whether the bond is stretched or compressed. The harmonic approximation means that the bond energy profile is estimated correctly only close to the equilibrium bond length and angle. The dihedral term consists of V_n referring to the height of the rotational energy barrier, n the periodicity of rotation, the dihedral angle ϕ, and the phase factor γ which allows the optimum dihedral angle to be offset from zero. The Lennard-Jones potential is commonly used to describe van der Waals interactions, where A_{ij} and B_{ij}, are empirical parameters for a given pair of

atoms, define the well depth and position of the minimum on the energy surface and R_{ij} is the interatomic distance. Finally, the electrostatic term, computed based on Coulomb's law, is composed of two partial atomic charges q_i and q_j with the distance R_{ij} between them and the dielectric constant ϵ.

By deriving the equation 1.1 with respect to changes in coordinates, one can get the initial forces on the atoms. The property of E_{MM} relates most closely to the thermodynamic quantity of the system internal energy (U) but it is difficult to relate between the zero points of E_{MM} and U. The difference in energies between any two states of the structure (ΔE_{MM}), in practice, is consequently used rather than the absolute energy.

1.3.2 Molecular dynamics simulations

Prior to performing molecular dynamics (MD) simulations, a technique known as energy minimisation is required to find an optimum molecular geometry representing a local energy minimum. Energy minimisation prevents the molecular system being fractured during MD simulations due to the large initial forces in the system. Although many minimisation algorithms are available, the steepest descent and conjugate gradient methods are widely used.

Each method has a different way to deal with the gradient of the energy as well as in its robustness and search efficiency. The steepest descent procedure is a powerful algorithm when the initial conformation is considerably different from the local minimum, whereas the conjugate gradient is more efficient when the structure is close to the minimum. Results from the minimisation step grant molecular conformations within a local minimum region that is close to the starting conformation. The ideal minimised conformation, in fact, is the one in the global energy minimum of the potential energy surface (PES). It is feasible to perform a systematic search of the conformational space by varying all torsion angles through the range of values then to determine the global minimum conformation. However, for a biological molecule systematic searching is not practical

as the number of searching conformations increases exponentially with the size of the molecule.

Once the atomic positions are assigned, the temperature of the system is usually elevated to 300 K leading to gain the different initial atomic velocities depending on the atomic mass and potential energy. Dynamic behaviour of the molecular system such as the fluctuations and conformational changes as a function of time can be calculated using MD simulations. The MD simulation is a conformation space search method in which the atoms with an initial velocity are allowed to move along time according to Newton's second law of motion [61],

$$F = ma \tag{1.2}$$

where F is the force exerted on the atom, m is the mass and a is its acceleration. The force applied on an atom can be determined by the change of the potential energy function (E_{MM}) and its coordinates r.

$$F = -\frac{dE_{MM}}{dr} \tag{1.3}$$

The total force on the conformation at time t from equation 1.3 is used to calculate the acceleration at time t in the motion equation 1.2. The Verlet algorithm (equation 1.4) [62] is a widely used method for integrating Newton's laws of motion in a MD simulation. With a certain time step δt (typically 1 or 2 fs), the Verlet algorithm uses the coordinates $r(t)$ and acceleration $a(t)$ at time t and the coordinates from the previous step, $r(t - \delta t)$ to calculate the new set of coordinates at time $t + \delta t$, $r(t + \delta t)$. During the time step, the exerted force on each atom is assumed to be constant. In every defined time step, equations 1.2 through to 1.4 are recalculated with the new positions of atoms and total forces, leading to a series of coordinate sets in space and time known as a trajectory.

$$r(t + \delta t) = 2r(t) - r(t - \delta t) + \delta t^2 a(t) \tag{1.4}$$

1.4 General methods

All MD simulations in chapter 2 were performed using AMBER suite version 8 [63] where simulations in chapter 3 and 4 were carried out using AMBER version 9 [64]. A recent review on the AMBER suite of programs has been published [65]. The AMBER ff03 force field [66] is a principal force field used in this work. Molecular visualization and manipulation were done using Insight II from Accelrys Inc., VMD - Visual Molecular Dynamics software [67] from the Theoretical and Computational Biophysics Group, NIH Resource for Macromolecular Modeling and Bioinformatics at the University of Illinois at Urbana-Champaign, and UCSF Chimera package [68] from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081).

1.4.1 System preparation

The starting coordinates of intrahelical FapydG and the FapydG-containing DNA-protein complex were taken from the PDB. An initial canonical *B*-DNA of a 12-mer oligonucleotide was generated by the *nucgen* module from the AMBER 8 package. Each model had a distinct method for preparation before solvation and will be separately described in the following chapters relating to the individual simulations.

All models were then solvated with an explicit water model, TIP3P [69], in a truncated octahedral box with a minimum 10 Å distance between any solute atom and a box edge. Periodic boundary conditions (PBC) were also applied to the systems for avoiding the impacts from the atoms on the outer surfaces drifting away form the simulation box. By means of PBC, the simulation system is surrounded with its images in three dimensions. Once an atom moves out of the box, an image atom moves in to replace it. The systems were also neutralised with an adequate number of potassium ions. In this study, potassium ions were preferred to sodium ions since it is evident that sodium ions highly

occupied the minor groove site and distorted the groove width in MD simulations where in lesser amount by potassium ions [70]. Solvation and neutralisation were performed using the *Leap* module; topology and coordinate files were finally generated.

1.4.2 Simulation conditions

Simulations were initially minimised by the short steepest descent technique and followed by the conjugate gradient method. Equilibration was performed using our multi-step protocol for gentle heating and harmonic restraint reduction over a period of 90 ps [71] and an 100-ps unrestrained equilibration was subsequently done. A production phase of the unrestrained MD was carried out at constant temperature (300 K) and pressure (1 atm) using the Berendsen algorithm [72] to control the simulation temperature by re-adjusting the atomic velocities. An integration time step of 2 fs was used and all bond lengths involving hydrogen were constrained using SHAKE [73]. The particle-mesh-Ewald (PME) approach [74] and PBC were used to estimate long-range electrostatic interactions with a nonbonded cutoff of 9 Å. Trajectories, which contain coordinates and velocities of each atom with respect to time, were saved every 2 ps over the course of production MD for further analyses. All simulations were carried out using the *sander* module as described above unless specified otherwise.

1.4.3 Basic trajectory analysis

Trajectory files were stripped of water molecules and potassium ions to keep solely common atoms of biomolecules. The stripped trajectory was analysed using a module called *ptraj* from the AMBER package. The module was used to generate some molecular parameter time-series throughout the trajectory such as bond lengths, bond angles and torsion angles. The time-averaged structure was also obtained.

In addition, a statistical method to calculate how much the conformation

changes along the time from some reference structure is the root mean square deviation (RMSD). The RMSD is calculated by

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} {D_i}^2} \tag{1.5}$$

where N is the number of atomic coordinates in each structure and D_i is the distance between the coordinates of atom i when the two structures are superimposed. A major advantage of RMSD values is that the RMSD plateauing may reflect the attainment of conformational stability.

The relative motion of each atom is calculated by the root mean square fluctuation (RMSF). The RMSF provides an information of atomic positional fluctuations around the average position $\bar{r_j}$ and may be calculated by

$$\text{RMSF} = \frac{1}{T}\sum_{i=1}^{T}\sqrt{\frac{1}{n}\sum_{j=1}^{n}(r_{i,j} - \bar{r_j})^2} \tag{1.6}$$

where T is the total number of frames in the trajectory, n is the number of atomic coordinates, and $r_{i,j}$ is the atomic coordinate in each frame.

1.5 Aims and objectives

With an incredible rate of up to 1 million DNA damage events per cell per day, cellular responses to DNA damage require a very powerful repair machinery to cope with such massive numbers and variations of damage. A primary repair mechanism is the BER pathway that is initiated by the lesion recognition by DNA glycosylases. Principal questions in the BER process are how the glycosylase enzymes find the damaged base in the DNA track and how they distinguish between damaged and undamaged bases. One distinct mechanism for target search is the protein sliding along DNA. Human oxoguanine DNA glycosylase 1 (hOGG1), for example, translocates along DNA with a rate of 1000 base pairs per 0.1 second and locates the lesion by a redundant search in which the enzyme

forms a specific complex with the lesion with a good time to flip and excise the damage from DNA [75]. Such a slow and redundant searching process, however, may be unable to handle the rapid production rate of DNA damage effectively. Thus, additional elements may contribute in the searching process to facilitate the repair protein.

The ultimate aims of this study were to gather information of dynamic processes of DNA damage recognition by DNA glycosylases and to reveal how an repair enzyme can find and discriminate between damaged and undamaged base using MD simulations. FapydG was selected for study because of its high mutagenicity and its difficulty to synthesise and be incorporated into a DNA duplex. Additionally, it was hypothesised that, in term of microscopic levels, the recognition process may be enhanced by some initial molecular signals from the damage such as the DNA flexibility and the distinct chemical structure of FapydG. Unfortunately, MD simulations are unable to handle with millisecond intervals to simulate the whole processes of recognition. MD simulations therefore have been employed to study three established areas of the lesion recognition process: bending, binding and flipping. To tackle the bending subject, unrestrained MD simulations of 12-mer oligonucleotides in the presence and absence of the lesion were performed. Stability and flexibility of FapydG-containing duplexes were mainly analysed compared to their normal equivalent. Regarding to the binding stage, DNA containing an extrahelical guanosine bound to Fpg was modelled based on the structure of the damaged DNA-Fpg crystal complex. Chemical interactions between the flipped-out FapydG and active residues in the binding pocket were determined compared to the normal nucleobase over the course of simulations. Finally, free-energy profiles of damaged and undamaged base flipping were generated in three situations: *B*-DNA, the distorted DNA structure and the protein-DNA complex.

Chapter 2

DNA Flexibility

2.1 Introduction

2.1.1 DNA flexibility and recognition

Protein-DNA recognition is an essential mechanism in many cellular processes, including in DNA damage repair, DNA replication and gene expression. The transcription process of genetic information from DNA or RNA and the restriction modification system require DNA-binding proteins that must accurately recognise specific sequences of DNA (e.g. a transcription factor or restriction endonuclease). Yet in other processes such as DNA repair, repair proteins must identify and correct damage to the DNA regardless of the DNA sequence context.

Generally, the ability of a protein to recognise its target DNA efficiently and specifically depends on two components: (i) the flexibility of DNA and protein, permitting them to adopt their optimal shape complementarity [76, 77, 78] and (ii) the contacts between amino acid sidechains and nucleotides form either direct hydrogen bonding, van der Waals and electrostatic interactions or indirect interactions through water molecules [79, 80]. It has been well documented that DNA frequently exhibits a bent structure when bound to a protein, whereas regular *B*-DNA is usually thought of as a straight double helix in the absence of a protein. However, many experiments have shown that double-helical *B*-DNA can be

bent without the protein and this intrinsic DNA bending is a sequence-dependent property [81, 82, 83, 84]. For instance, adenine tracts (A-tracts) with four to six consecutive adenine nucleobases have shown DNA bending of $\approx 18°$ [85]; DNA bending of 23° was observed at the central GC step in the crystal structures of two different decamers, $d(CATGGCCATG)_2$ and $d(CCAAGCTTGG)_2$ [82, 83].

Recognition of the damaged base by an appropriate repair enzyme is the first step in DNA repair processes. Crystallographic studies of damaged DNA-glycosylase complexes showing a bent DNA structure with an extrahelical lesion when bound to a glycosylase suggest a key feature in the BER recognition. As shown in table 2.1, bend angles of lesion-containing DNA in the complexes are to a much greater degree than those occur in the intrinsically curved DNA. Hence it was hypothesised that base damage could be one causative factor of such high degree of bending.

Table 2.1: Bend angles of lesion-containing DNA bound to glycosylase enzymes (analysed by CURVES program [86]).

Protein	DNA lesion	PDB ID	Bend angle (°)	Reference
Fpg	FapydG	1XC8	63.5	[54]
hOGG1	8OG	1EBM	64.0	[87]
AlkA	1-Azaribose	1DIZ	66.0	[88]
AAG	EDA	1EWN	27.2	[89]
UDG	Uracil	1SSP	38.1	[90]

Abbreviations: Fpg, formamidopyrimidine-DNA glycosylase; FapydG, 2,6-diamino-4-hydroxy-5-formamidopyrimidine; hOGG1, human 8-oxoguanine glycosylase; 8OG, 8-oxoguanine; AlkA, *E.coli* 3-methyladenine-DNA glycosylase II; AAG, human 3-methylalkyladenine-DNA glycosylase; EDA, $1,N^6$-ethanoadenine; UDG, uracil-DNA glycosylase

Unfortunately, there is little information regarding the microscopic effects of DNA lesions on conformational flexibility. But studies on the backbone structure of a 25-base DNA strand containing single 8-oxopurine using Fourier transform-infrared spectroscopy with multivariate statistics demonstrated that the backbone deformation likely occurred immediately adjacent to the lesion [91]. MD simulation studies of normal and 8OG-containing DNA showed that bending dynamics

towards the major groove was significantly more likely in the presence of the lesion [92]. These findings suggest that DNA deformation is crucial in the initial recognition mechanism.

2.1.2 Structural analysis of the FapydG residue

The opened imidazole ring of the FapydG residue leads to an increased degree of freedom in the glycosidic bond and the formamide group. The crystal structure of Fpg bound to an extrahelical *c*FapydG (PDB ID 1XC8) [54] as well as the NMR structure of an intrahelical formamidopyrimidine adduct of aflatoxin B1 (PDB entry 1HM1) [93] revealed that the FapydG presents the *anti*-glycosidic conformation with a nonplanar *cis*-formamido group (*anti-cis*-FapydG). On the other hand, an experimental study of FapydG on DNA polymerase activity showed that FapydG induces misincorporation of adenine opposite FapydG as the lesion adopts a *syn*-conformation and the FapydG:A base pair presents a thymine-like hydrogen bonding pattern [47]. Thus, it is certainly important to first determine the reasonably intrahelical conformation of FapydG before incorporating the FapydG residue into DNA models.

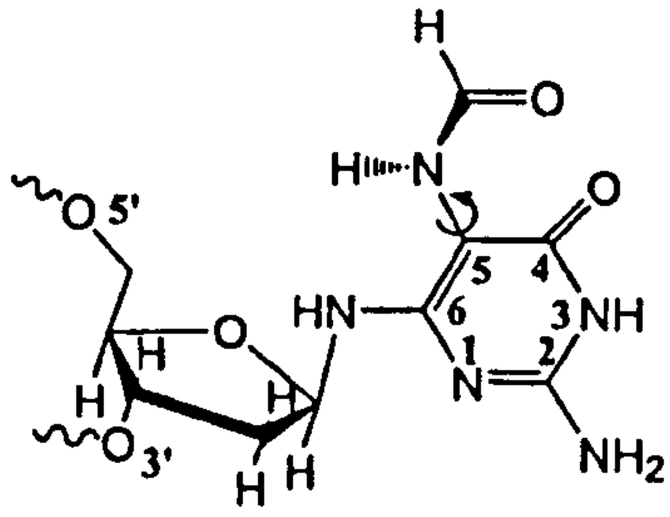


Figure 2.1: The structure of *anti-cis*-FapydG presenting *anti*-glycosidic bond and the rotatable C5-N5 bond in a *cis*-configuration.

2.1.3 Aims and objectives

The aim of the work in this chapter was to identify the initial molecular perturbation to the normal structure of duplex DNA brought about by the presence of FapydG, which may trigger the damage recognition by Fpg. The influence of

FapydG on DNA stability and flexibility compared to its normal equivalent were studied using an MD simulation technique. Prior to performing MD simulations, a potential structure of an intrahelical FapydG with an appropriate force field parameter set was first developed based on both the X-ray crystal and the NMR structures available to date plus theoretical methods. Unrestrained MD simulations were performed on a dodecamer DNA duplex containing a single *syn*- or *anti*-FapydG residue paired with cytosine, to determine which conformation of the lesion was more energetically favourable. Further dynamic simulations on a series of DNA duplexes containing FapydG:C or G:C were carried out to study the stability and flexibility of the duplexes. Subsequently, global axis curvature was calculated to investigate the intrinsic curvature of FapydG-containing DNA; the resulting curvature analysis was compared to that observed in crystal structures of damaged DNA bound to Fpg. Finally, the energetic penalty for different DNA sequences to adopt the observed protein-bound conformation were estimated by a combination of the principal component analysis and a novel approach based on the calculation of a statistical measure, the Mahalanobis distance.

2.2 Simulation methods

2.2.1 FapydG parameterisation

Since FapydG is a non-standard nucleotide, there is no available molecular parameters included in the database as is the case for standard DNA and RNA nucleotides. In order to incorporate a FapydG base into the simulations, it was first essential to assign an appropriate type and partial charge to each atom. A methyl derivative of FapydG was modelled, in which the deoxyribose moiety was substituted by a methyl group keeping the FapydG in a neutral system. Electrostatic potential on the molecular surface was calculated using Gaussian 98 [94], with Hartree-Fock calculations and the 6-31G* basis set. To reproduce this electrostatic potential, appropriate partial charges were then fitted at each atom using

an atom-centred point charge model termed the restrained electrostatic potential (RESP) method [95] within the *antechamber* module. The charges were finally merged with standard AMBER charges for nucleic acid sugars and phosphates, and adjusted to give the correct total overall charge. Atom types of FapydG were assigned based on the standard AMBER parameters for organic and biomolecular molecules, parm99.dat [96].

The FapydG force field was calculated according to the parm99.dat with the missing torsional terms for X-C5-N5-X (see figure 2.1). To analyse rotational energies around the C5-N5 bond, snapshots were generated ranging 0° to 180° with 30° increments using Insight II. A full geometry optimisation and energy calculation using the 6-31G* basis set with each torsion angle constrained was performed. The AMBER energy profile of the dihedral angles around the C5-N5 bond was also calculated and subsequently fitted with the quantum mechanically potential energy profile using different torsional parameters. Standard atom types and charges were used for the remaining nucleic acid residues.

2.2.2 Model construction and simulations

2.2.2.1 Glycosidic bond conformations

The potential conformation about the glycosidic bond of intrahelical FapydG was first examined. An initial 12-mer *B*-DNA duplex d(CTTTTGCAAAAG)$_2$, termed TGC, was generated by the *nucgen* module [63]. The dynamics of the TGC sequence was extensively studied in previous work [97, 98]. It was shown that the sequence remained stable over the 5-ns simulations and its time-averaged structure was in a great agreement with the NMR-derived structures. The parameterised FapydG was then incorporated into the TGC sequence to give d(CTTTTFCAAAAG)·d(CTTTTGCAAAAG) termed TFC where F was FapydG; an intrahelical *anti* conformation (240°) of FapydG was employed from the NMR structure of the formamidopyrimidine adduct of aflatoxin B1 [93]. A sample of DNA containing *syn*-FapydG was also remodelled based on a *syn*-

deoxyguanosine giving a glycosidic angle of 74° after minimisation and equilibration. Both conformations were paired with cytosine as shown in figure 2.2. Each system was solvated, equilibrated and simulated for 5 ns as described in section 1.4.2. The simulations were analysed to identify the most energetically favourable glycosidic conformation of FapydG prior to carrying out further experiments.

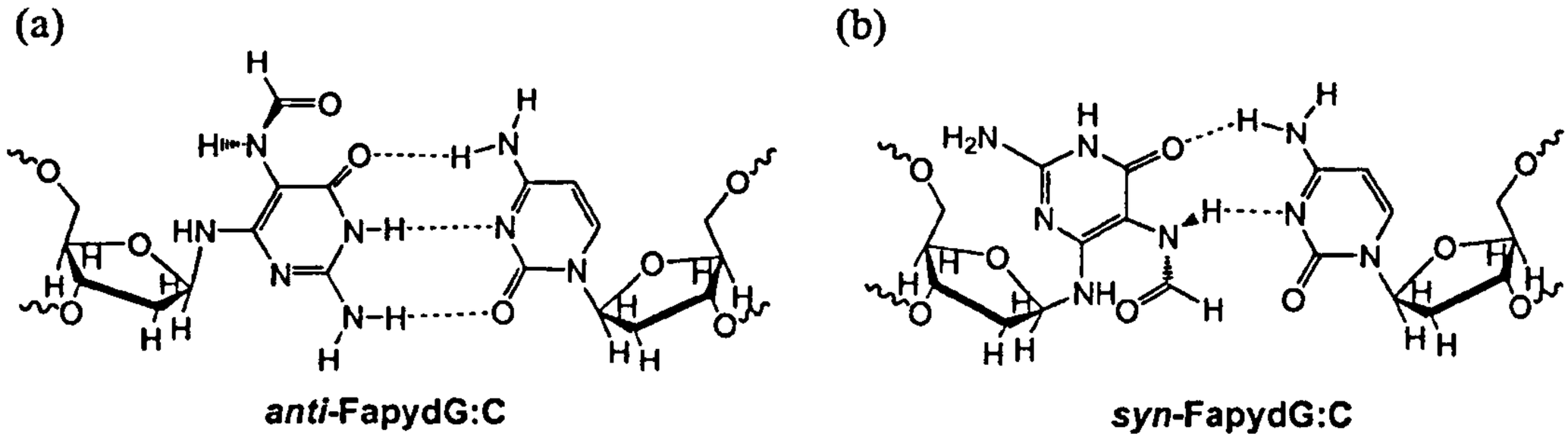


Figure 2.2: Starting base pairing of (a) *anti*-FapydG:C with a G:C-like hydrogen-bonding pattern and (b) *syn*-FapydG:C with an A:T-like pattern.

2.2.2.2 Extended simulations for sequence-dependent flexibility

To check if the conformational preferences of the FapydG base were influenced by its neighbouring bases, the original model systems, featuring a TXC sequence (X is G or FapydG) were modified (using Insight II) to give AXA, AXC, TXA, AXG and TXT-containing sequences. The 5-ns simulations were carried out as described in section 1.4.2.

2.2.2.3 Conformational search for the target structure

Coordinates of the bent DNA d(TCTTTFTTTTCTC)·d(GAGAAACAAAGA) containing FapydG were initially obtained from the PDB (PDB code 1XC8, [54]). This distorted DNA conformation from the x-ray crystallographic structure was then defined as the 'target' structure. To allow for the fact that the target conformation is probably somewhat flexible, and the crystal structure could not represent this, the simulation technique was applied in order to produce a set of target conformations. The DNA structure was solvated, minimised and equili-

brated as described in section 1.4.2. A short restrained molecular dynamics with 5 $kcal/mol/Å^2$ cartesian restraint on both terminal base pairs was then carried out to sample its possible deformed conformations of DNA during the simulation. Coordinates were saved every 1 ps of the 100 ps restrained simulation. These coordinates were analysed using CURVES 5.1 [86] to collect the angle and magnitude of global axis curvature as a reference ensemble of target conformations.

2.3 Post simulation analysis

2.3.1 Principal component analysis

Principal component analysis (PCA) or essential dynamics (ED) is one of the major tools for the analysis of biomolecular MD simulation [99, 100]. By the means of PCA, an output trajectory which contains a mass of numerical data can be analysed objectively and reduced in complexity to identify the most important dynamical behaviour and calculate their relative significance in the overall motion of DNA. The *ptraj* module in AMBER 8 suite [63] was employed to calculate the principal components from trajectories.

Prior to performing PCA, the RMSD of snapshots from the initial configuration and from the time-averaged one was calculated over all DNA atoms to remove the transitional and rotational degrees of freedom, and to produce a best fit of the coordinates of the entire DNA. Each 5-ns trajectory was used to construct the positional covariance matrix with $3N \times 3N$ elements, where N was the number of atoms in the system ($N = 762$ or 763 for G and $N = 765$ or 766 for FapydG). Elements c_{ij} in the Cartesian covariance matrix $\mathbf{C}$ are computed by

$$c_{ij} = \frac{1}{M} \sum_{t=1}^{M} [(r_i(t) - \bar{r}_i) \cdot (r_j(t) - \bar{r}_j)] \tag{2.1}$$

where M is the total number of snapshots, and r_i and r_j denote a given pair of coordinates [101]. Since $\mathbf{C}$ is a square matrix, diagonalisation of the $\mathbf{C}$ matrix yields a set of $3N - 6$ eigenvectors and their associated eigenvalues, in which

contains the same fundamental properties of the square matrix, as follows

$$\mathbf{C} = \mathbf{VDV^{-1}} \tag{2.2}$$

where $\mathbf{V}$ denotes a matrix composed of the eigenvectors ($\mathbf{v_n}$) of the covariance matrix $\mathbf{C}$, $\mathbf{D}$ is the diagonal matrix constructed from the corresponding eigenvalues (λ_n), and $\mathbf{V^{-1}}$ is the matrix inverse of $\mathbf{V}$. The eigenvector describes a vectorial representation of each mode of structural deformation that may be expressed as a part of the total motion, and the eigenvalue for a mode indicates the relative contribution that this mode has made to overall motion within the trajectory [102]. The larger number the eigenvalue the more the contribution of the motion, thus only those with high eigenvalues are usually considered.

PCA simplifies multidimensional data into scalar data which re-express the original information by projection of the trajectory $\mathbf{r(t)}$ onto individual eigenvectors ($\mathbf{v_n}$) giving their coefficients ($p_n(t)$) in each snapshot over the time (see equation 2.3). Coefficients or projections for a particular eigenvector imply the mode of DNA deformation in the eigenvector subspace. The probability distributions of projections were analysed to ensure the equilibration of the system. Artificial animations were generated from its maximum and minimum projection values for major modes of DNA deformation.

$$p_n(t) = \mathbf{v_n} \cdot \mathbf{r(t)} \tag{2.3}$$

Furthermore, a set of major eigenvectors from different molecular system i.e. the modified and unmodified DNA duplexes were quantitatively compared using the dot product,

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cdot \mathbf{cos}\theta \tag{2.4}$$

where $|\mathbf{a}|$ and $|\mathbf{b}|$ are the magnitude of eigenvectors $\mathbf{a}$ and $\mathbf{b}$, and θ is the angle between those two vectors. The dot product provides the degree of similarity of DNA motion between two selected eigenvectors.

2.3.2 Energetic analysis

2.3.2.1 Relative binding energies

Coordinates every 50 ps were analysed by the MM/GBSA approach to calculate relative free energies of binding. The MM/GBSA method combines molecular mechanics, a continuum implicit solvent model termed a generalized Born (GB) model, and a solvent accessibility (SA) method to estimate free energies of molecules (G_x) in solution (see equation 2.5). With the GB solvent model, estimating solvation free energies becomes feasible because solvent degrees of freedom are taken into account implicitly [103, 104]. The molecular mechanical energies: the van der Waals interaction (E_{vdW}), electrostatic interaction (E_{es}), and internal energy (E_{in}), as well as the free energy of GB solvated systems (G_{pol}) are calculated with the *sander* program [63]. The hydrophobic contribution to the solvation free energy (G_{nonpol}) is estimated using the solvent accessible surface area of the solute [105].

$$G_x = E_{vdW} + E_{es} + E_{in} + G_{pol} + G_{nonpol} \tag{2.5}$$

In this study, DNA (without FapydG or G) was treated as the "receptor" and the FapydG or guanine nucleobase as the "ligand". The G_{pol} was calculated using GB solvent model developed by Onufriev, Basford and Case (GB^{OBC}, igb=5) [106] as recommended by Kormos and Beveridge for DNA simulations [107], where the G_{nonpol} was calculated from equation 2.6 [108]. The relative binding energy ($\Delta G_{binding}$) represents the contribution of the nucleobase-DNA interactions to the relative stability of these two structures during the simulations. The relative binding free energy is estimated with equation 2.5 through 2.8.

$$G_{nonpol} = 0.0072\ kcal/mol/\mathring{A}^2 \times\ SA \tag{2.6}$$

$$\Delta G_{binding} = G_{complex} - G_{receptor} - G_{ligand} \tag{2.7}$$

$$\Delta G_{binding} = \Delta E_{vdW} + \Delta E_{es} + \Delta E_{in} + \Delta G_{pol} + \Delta G_{nonpol} \quad (2.8)$$

2.3.2.2 Intra- and inter-strand interactions

Intra- and inter-strand interactions were calculated over each 5-ns simulation using the *anal* module of AMBER 7 [109] to represent stacking interactions and base pair hydrogen bonding respectively. In all cases the backbones were removed, the charge on C1' was adjusted to have a total neutral charge in each base and van der Waals parameters were applied for all atoms except for C1'. All non-bonded interactions were determined for 3' and 5' bases adjacent and the opposite cytosine to the lesion or the central guanine.

2.3.3 Global bending analysis

Helical bending occurs when there are local distortions of the helix geometry. The local distortions, however, do not certainly result in a global bending because those local deformations may neutralise each other giving a normal straight helix. Thus only global bending parameters are of interest in this study. To analyse global axis curvature, a well-known program called CURVES version 5.1 [86] was employed. One advantage of the CURVES method over other approaches like FREEHELIX [110] is that it can calculate global helical axis parameters and then optimise them to yield the best curvilinear helical axis [111]. After the optimisation, the angular magnitude is then calculated from the angle between the first and the last helical axis segments of the curvilinear helical axis and the direction is the angle of the helical axis with respect to the dyad axis (pointing towards the minor groove) of the first nucleotide (figure 2.3). The terminal base pairs were omitted from all simulations prior to performing the analysis due to the fraying effects. The global curvature parameters were determined and compared with those from the crystal structure and the short restrained simulation.

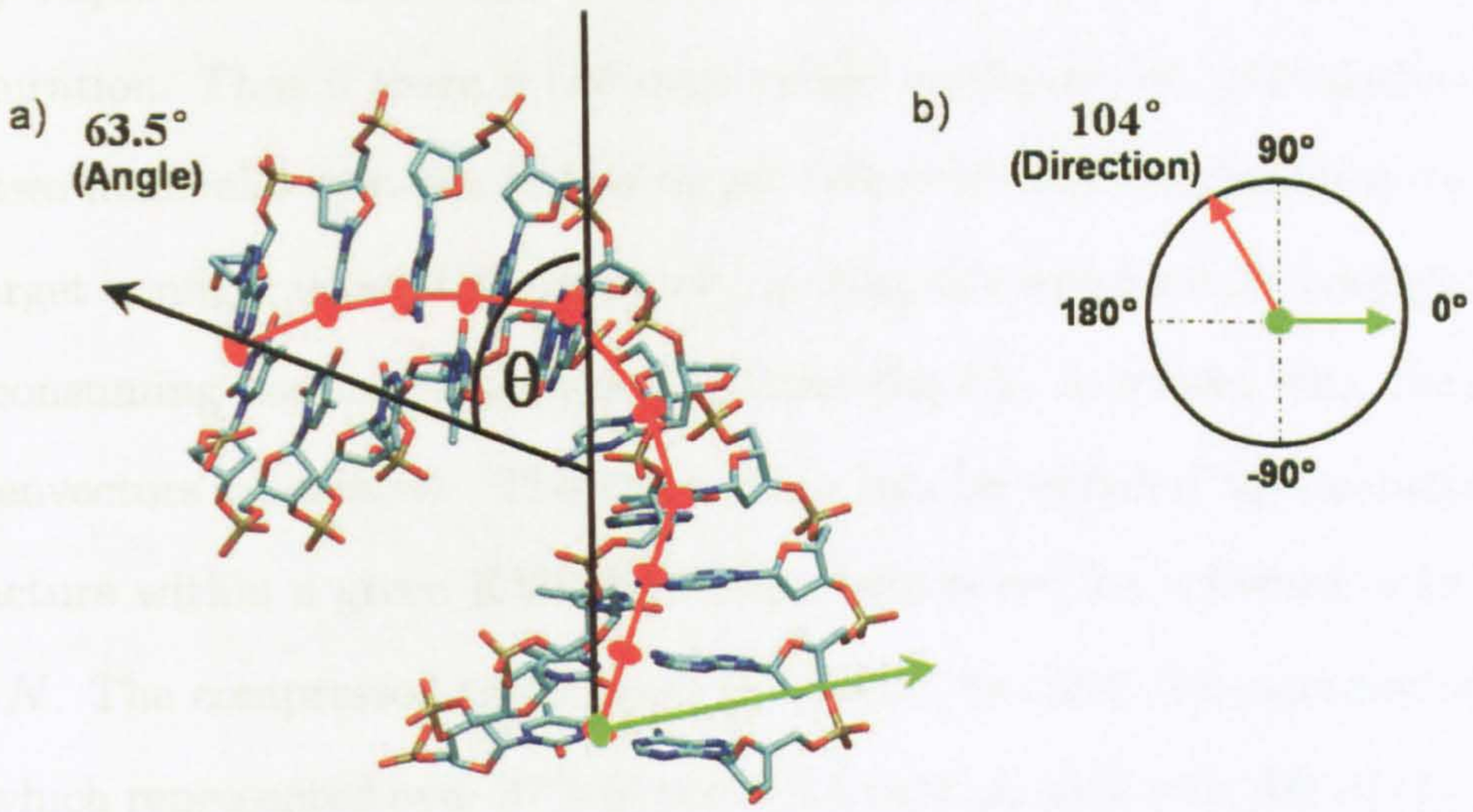


Figure 2.3: A schematic picture for calculations of angular magnitude and direction by CURVES 5.1 [86] (a) the curvature angle (θ) formed between the first and the last helical axes (black lines) of the curvilinear axis (red line) calculated from each helical axis segment (red circle) (b) the angular direction (red arrow) estimated with respect to the dyad axis of the first base (green arrow).

2.3.4 Mahalanobis distances

The Mahalanobis distance (D_M) is a distance measure between an unknown data set and a known one by taking into account the correlations of the data set [112]. D_M calculations were employed to measure the distance of any configuration from the average structure using the PCA data. The D_M was calculated by

$$D_M = \sqrt{\sum_{i=1}^{N} \frac{p_i^2}{\lambda_i}} \tag{2.9}$$

where p_i and λ_i are the projection and eigenvalue along each principal component, respectively, and N is the number of eigenvectors. Since all translational and rotational modes are removed from the trajectory before performing PCA, the resulting eigenvectors are assumed to be equal to normal vibrational modes of the system and their corresponding eigenvalues can refer to the forces associated with DNA deformation along the modes. Up to this point, the D_M relates to the

energy required to distort the structure from the average configuration to any configuration. Thus if there is the same target configuration, calculations of D_M from two molecular systems to this target reflect to their attainability to achieve the target configuration. Unfortunately, getting the exact target configuration is time-consuming and overestimates D_M since the D_M increases with the number of eigenvectors calculated. The calculation can be speeded up by targeting to a structure within a given RMSD tolerance which can be achieved with a much lower N. The compressed trajectories [113] with the first 100 eigenvectors ($N = 100$) which represented over 97% of the DNA motion with only 5% of the original file size were considered. Ten central base-pairs were used to calculate the D_M in order to avoid the fraying effects.

The concept of D_M measurement was applied to the compressed trajectories to analyse DNA deformability of modified and unmodified duplexes towards the conformation required for protein binding within a conformational space of RMSDs of 2.0 and 1.0 Å. The search method to find the combination of individual projections along each eigenvector that minimises the D_M from the original (time-averaged) structure to the target (crystal) one within a given RMSD tolerance was introduced as shown in figure 2.4. The initial step was to calculate the dot product vector between those two structures and then decide the direction to deform the original structure. The next step was that the starting structure is distorted by moving a short distance along each eigenvector in turn. After each move, the change (reduction) in RMSD, and the change (increase) in D_M were calculated. When the eigenvector that results in the biggest reduction in RMSD for the smallest increase in D_M was found, the structure was subsequently moved to that new position and became the new starting structure for the next cycle of tests of different moves. The search process was continued repeatedly and stopped when either (a) the minimum RMSD was reached or b) a maximum D_M was reached or c) no move could be found that lowers the RMSD any further.

According to the harmonic approximation, a D_M can be transformed into an energetic penalty required for deformation between two well-defined conformations using a Hooke model (see equation 2.10)

$$E_{def} = \frac{k_B T}{2} D_M^2 \tag{2.10}$$

where E_{def} denotes the deformation energy, k_B is the Boltzmann constant and T is the simulating temperature.

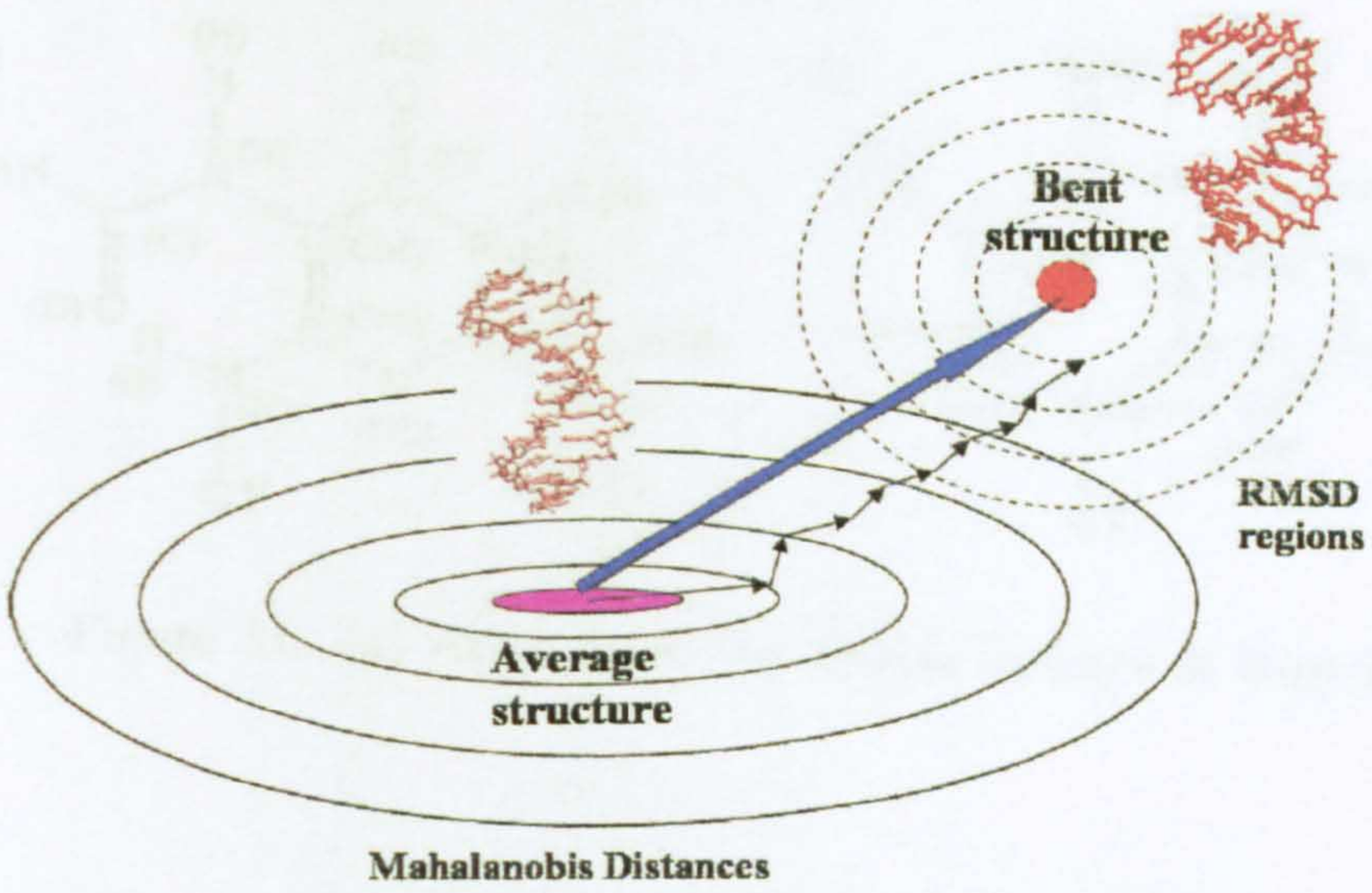


Figure 2.4: A minimal Mahalanobis distance searching algorithm from an average structure to a target conformation within the RMSD tolerance. The blue arrow shows the principal direction to deform the original structure and each small black arrow represents the deformation of the structure to achieve a minimal RMSD along each eigenvector.

2.4 Results and discussions

2.4.1 Modified AMBER force field for FapydG

Missing force field parameters for FapydG were first reported by Perlow-Poehnelt *et al.* [114]. The FapydG parameters have been mostly maintained similar to the guanine nucleobase leading to the constrained formamide group in the plane with

the pyrimidine ring. In contrast, a nonplanar formamide group was expected as occurred in the X-ray and the NMR structures.

Applying the new modified force field, preliminary MD simulations of a FapydG-containing oligonucleotide in the presence and absence of Fpg showed the FapydG conformation in good agreement with the crystallographic and the NMR structures. The resulting partial charges and atom type assignments are shown in figure 2.5 and the modified force field for FapydG used in this study is demonstrated in table 2.2.

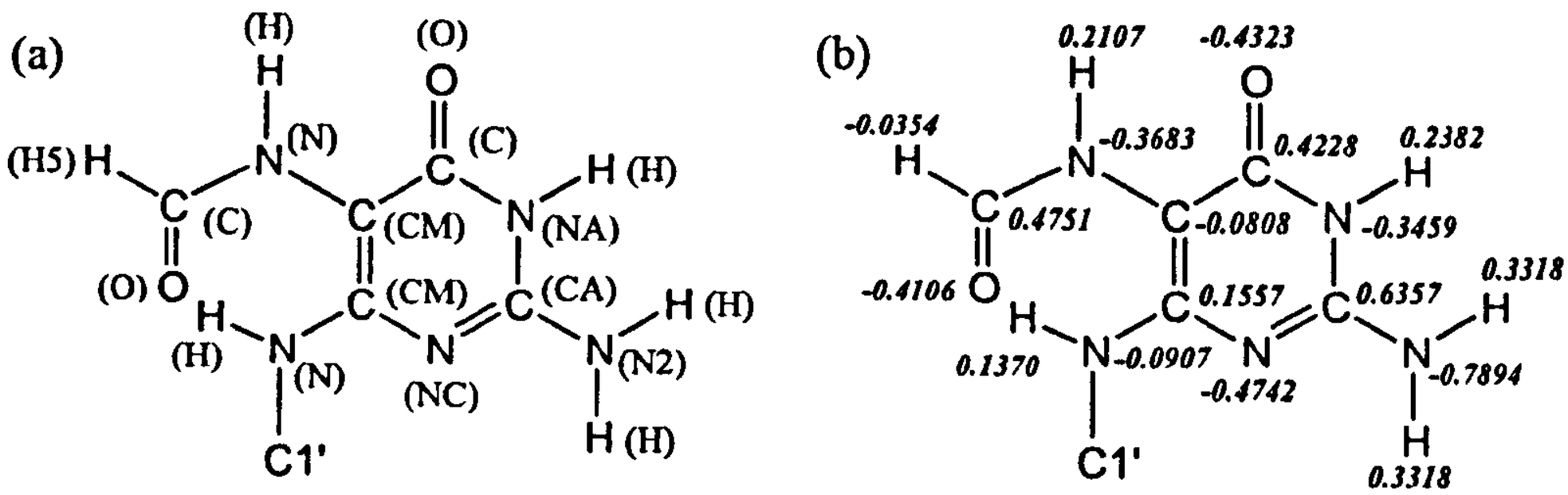


Figure 2.5: (a) Atom types (b) Atomic charges of FapydG.

2.4.2 Glycosidic conformation of FapydG

RMSDs of snapshots from the initial conformation over the entire DNA atoms for 5-ns MD trajectories of both *syn*- and *anti*-FapydG were first calculated to produce the least-square fit of the coordinates. As depicted in figure 2.6(a), the RMSD with the respective starting structure indicates that the *syn*- and *anti*-FapydG containing DNA systems are stable during the fully unrestrained simulations with average RMSD values of 1.92 and 1.75 Å respectively. The *ptraj* module [63] was then employed to track the glycosidic torsion angle (O4'-C1'-N6-C6) over the simulation. It has also been clearly evident in figure 2.6(b) that *syn* → *anti* conversion occurred after 500 ps and was maintained along the simulation.

Table 2.2: Modified AMBER parameters added to the parm99.dat.

Bond	K_{bond}	R_{bond}	Analogy with
N-CM	448.0	1.365	CM-N*
NC-CM	483.0	1.339	CA-NC
Bond angle	K_{angle}	R_{angle}	**Analogy with**
C-N-CM	70.0	121.60	C-N*-CM
N-CM-C	70.0	120.10	CM-CA-N2
N-CM-CM	70.0	121.20	CM-CM-N*
H-N-CM	50.0	121.20	CM-N*-H
CM-CM-NC	70.0	121.20	CM-CM-N*
CA-NC-CM	70.0	118.60	CA-NC-CQ
NC-CM-N	70.0	119.30	N2-CA-NC
N-CT-H2	50.0	109.50	N*-CT-H1
CT-N-CM	70.0	121.20	CM-N*-CT
OS-CT-N	50.0	109.50	OS-CT-N*

Dihedral	Phase	$K_{dihedral}$	Phase angle	Periodicity	Analogy with
CA-NC-CM-N	1	1.10	180.0	2.	CB-NC-CA-N2
CM-CM-NC-CA	1	1.85	180.0	2.	X-CM-N*-X
H-N-CM-NC	1	1.85	180.0	2.	X-CM-N*-X
CT-N-CM-NC	1	1.85	180.0	2.	X-CM-N*-X
CT-N-CM-CM	1	1.85	180.0	2.	X-CM-N*-X
H-N-CM-CM	1	0.25	0.0	2.	New parameter
H-N-CM-C	1	0.25	0.0	2.	New parameter
C-N-CM-CM	1	0.25	0.0	2.	New parameter
C-N-CM-C	1	0.25	0.0	2.	New parameter

Notes: K is a force constant for each term and R is an ideal value.

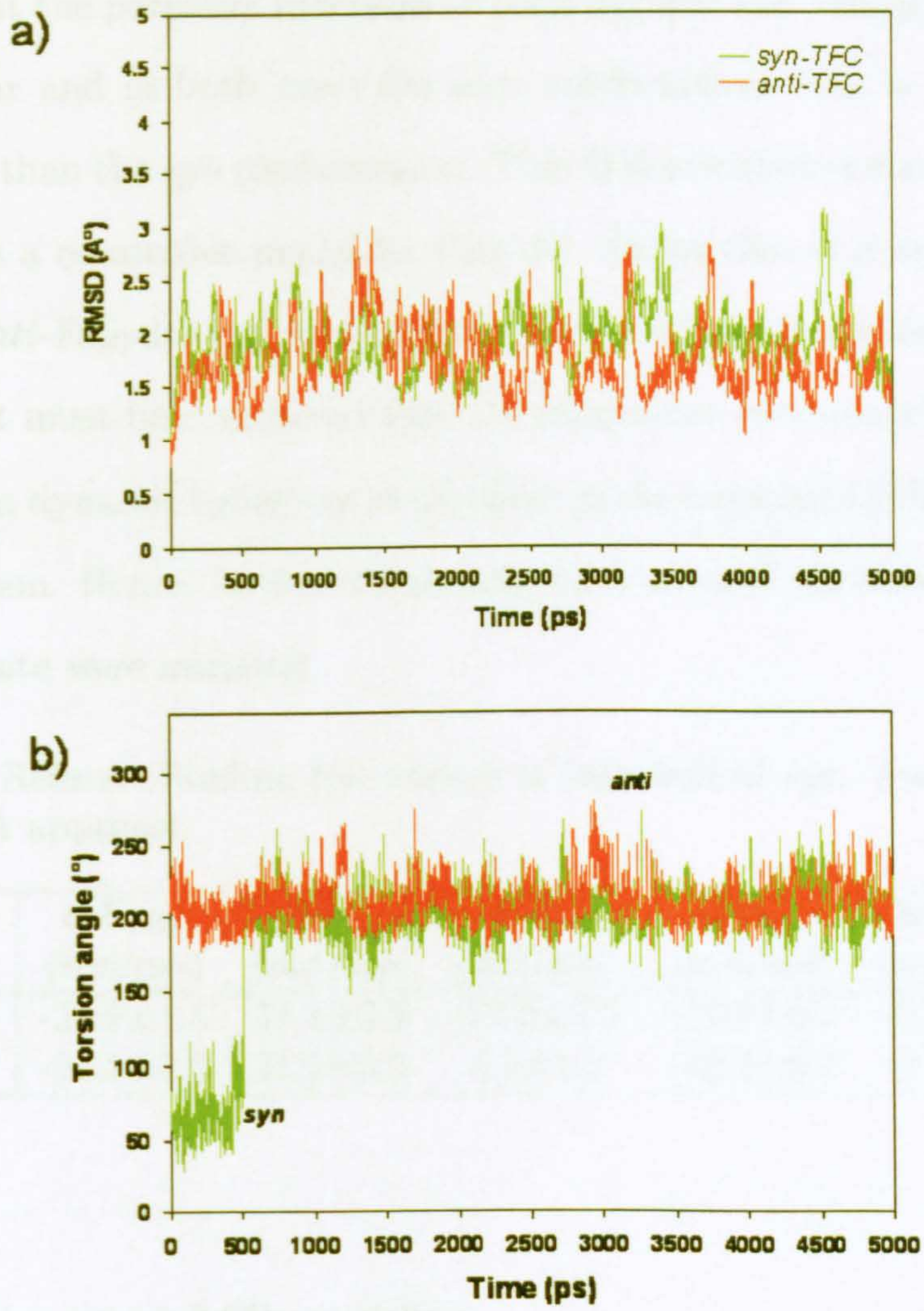


Figure 2.6: (a) RMSD plots of the *syn*- and *anti*-FapydG, compared to the starting structure as a function of simulation time (b) Glycosidic torsion angles of the *syn*- and *anti*-FapydG over the 5-ns simulations.

The MM/GBSA approach was subsequently used to calculate relative binding free energies of both conformations within the duplex as shown in table 2.3. The results of the two binding modes are roughly the same in any energetic term except for the electrostatic interaction. The *anti* conformation is more stabilised than the *syn* structure with $\approx 11\ kcal/mol$, since the *syn*-FapydG:C pairing exhibits a double hydrogen bond whilst a triple hydrogen bond is present in *anti*-FapydG:C pairing. This preliminary result here has a good agreement with the quantum mechanical calculation by Ober *et al.* [115]. The density functional calculation

with a constraint on the glycosidic torsion angle of either *c*FapydG or FapydG showed that the potential functions of both natural and analogue molecules were very similar and in both cases the *anti* conformation was ≈ 6 $kcal/mol$ more favourable than the *syn* conformation. This QM calculation also suggests that the *c*FapydG is a reasonable model for FapydG. So far then it is strongly convincing that the *anti*-FapydG is a potential structure for the intrahelical FapydG lesion. However, it must be considered that the nonplanar formamide group may cause the different dynamic behaviour depending on the adjacent DNA base particularly in 3' direction. Hence, further simulations on a series of nucleobases neighbouring the lesion site were initiated.

Table 2.3: Relative binding free energy of intrahelical *syn*- and *anti*-TFC from MM/GBSA approach

Sequence	ΔE_{vdW} (kcal/mol)	ΔE_{es} (kcal/mol)	ΔE_{in} (kcal/mol)	ΔG_{pol} (kcal/mol)	ΔG_{nonpol} (kcal/mol)	$\Delta G_{binding}$ (kcal/mol)
syn-TFC	-23.9±1.5	45.1±3.6	-12.4±3.5	-24.4±6.2	-2.5±0.1	-18.0±6.8
anti-TFC	-24.1±2.0	21.5±3.6	4.2±1.3	-28.2±6.2	-2.3±0.1	-28.9±5.4

2.4.3 General MD results

A nucleobase of FapydG or guanine paired with cytosine in Watson-Crick geometries was incorporated in the central AXA, AXC, AXG, TXA, TXC and TXT sequences. The RMSD time series of each duplex, which provides a measure of conformational changes with respect to the initial conformation, fluctuates more broadly in the presence of FapydG, as depicted in figure 2.7. The average RMSD for each sequence is shown in table 2.4. The RMSD generally implies that the lesion-containing DNA would have more conformational flexibility than its normal counterpart, however there are exceptions to this rule. No tendency of base opening was detected in all 5-ns simulations.

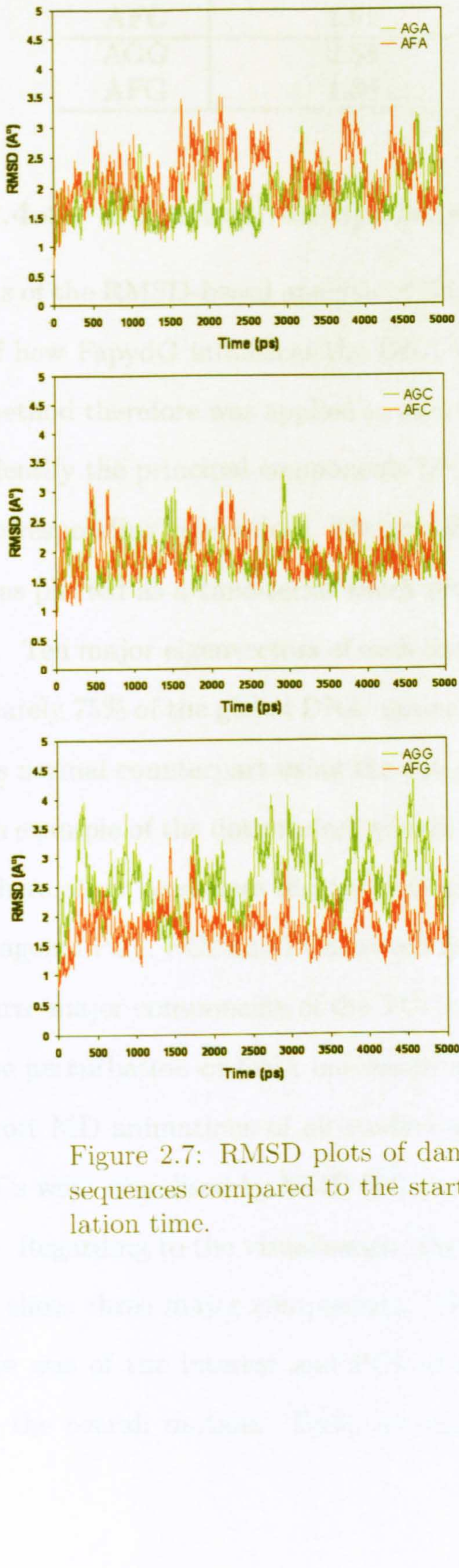


Figure 2.7: RMSD plots of damaged (red) and undamaged (green) sequences compared to the starting structure as a function of simulation time.

Table 2.4: Average RMSD values of damaged and undamaged sequences compared to the starting structure over its 5-ns simulation.

Sequence	Average RMSD (Å)	Sequence	Average RMSD (Å)
AGA	1.81	TGA	1.90
AFA	2.16	TFA	2.26
AGC	1.86	TGC	1.84
AFC	1.91	TFC	1.75
AGG	2.58	TGT	1.67
AFG	1.80	TFT	1.99

2.4.4 Principal component analysis

As of the RMSD-based analysis of MD trajectories, there is no obvious indication of how FapydG influences the DNA behaviour in terms of flexibility. The PCA method therefore was applied to each of twelve 5-ns trajectories independently to identify the principal components (PCs) of DNA motion (eigenvectors) and their corresponding eigenvalues. The projection of the trajectory onto each component was plotted as a time-series which reveals the global movement for each mode.

Ten major eigenvectors of each damaged duplex, which contribute to approximately 75% of the global DNA motion, were subsequently compared with those of its normal counterpart using the dot product approach. Figure 2.8 demonstrates an example of the dot product matrix of damaged and undamaged TGC sequence where other sequences display similar results. The matrix shows that along the diagonal PC1, PC2 and PC3 are red indicating a high degree of similarity between three major components of the TGC and TFC sequences. The matrix also shows the perturbation of DNA movement due to the presence of FapydG. Eventually, short MD animations of all studied sequences projected along first three major PCs were visualised by VMD [67] to observe the type of motion.

Regarding to the visualisation, the bending and twisting modes were observed in those three major components. The PC1 presenting a bend about the lesion site was of the interest and PC1 of all sequences contributed averaging 26.4% to the overall motion. Table 2.5 shows the proportion of first three principal

components that contributed to the global DNA movement. It was nonetheless not possible to obtain a clear picture of the impact of FapydG on DNA duplex by viewing animations. The PC1 motion was subsequently analysed by CURVES 5.1 and will be discussed in section 2.4.6.

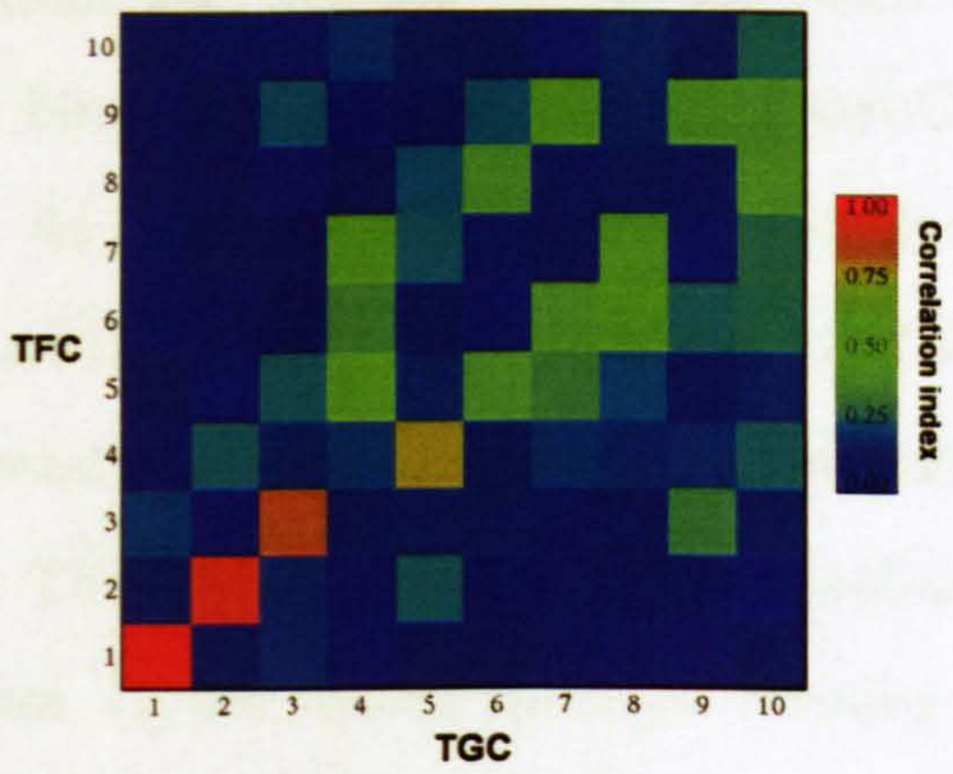


Figure 2.8: Dot product matrix for damaged and undamaged TGC DNA sequence, scaling from 0 to 1 which shows 1 to be a red square corresponding to perfectly overlap between two eigenvectors and 0 in a blue square indicating no overlap between the vectors.

Table 2.5: Proportion of first three principal components of each sequence contributed to the overall motion.

Sequence	PC1 (%)	PC2 (%)	PC3 (%)
AGA	24.5	17.2	12.1
AFA	24.7	17.4	11.8
AGC	20.8	16.7	12.0
AFC	36.5	15.8	8.9
AGG	25.0	18.4	10.3
AFG	23.5	17.1	12.6
TGA	24.3	16.7	8.3
TFA	24.2	18.7	10.9
TGC	27.9	16.1	7.7
TFC	25.8	17.8	12.2
TGT	29.3	12.6	9.3
TFT	30.5	16.4	6.8

2.4.5 Energetic properties and DNA stability

2.4.5.1 Relative binding energies

A low melting point of the *c*FapydG:C base pair compared to G:C base pairs in the duplex has been reported [53]. It suggests the dramatic destabilisation of the FapydG-containing duplexes. The MM/GBSA approach was initially performed to estimate relative binding free energies of the FapydG base compared to its normal counterpart. As shown in table 2.6, in all the cases $\Delta G_{binding}$ of damaged duplexes are less than the normal duplexes about 5-10 $kcal/mol$ mainly destabilised by the electrostatic interaction (ΔE_{es}) and the polar part of the solvation free energy (ΔG_{pol}). The destabilisation of DNA containing FapydG may be explained by two factors: (i) the weaker hydrogen bonding of the FapydG:C base pair due to the high degree of freedom of the glycosidic bond (ii) the negative charges of the nonplanar formamide group leading to unfavourable electrostatic interactions with the 3'-flanking base.

Table 2.6: Estimated relative binding free energy of damaged and undamaged DNA duplexes from MM/GBSA approach.

Sequence	ΔE_{vdW} (kcal/mol)	ΔE_{es} (kcal/mol)	ΔE_{in} (kcal/mol)	ΔG_{pol} (kcal/mol)	ΔG_{nonpol} (kcal/mol)	$\Delta G_{binding}$ (kcal/mol)
AGA	-24.0±2.1	17.6±4.5	6.5±1.2	-31.9±7.1	-1.8±0.1	-33.6±6.4
AFA	-23.6±2.1	23.1±3.6	4.1±1.2	-28.8±6.2	-2.1±0.1	-27.2±5.9
AGC	-24.5±2.6	18.1±4.4	6.9±1.2	-31.7±8.2	-1.9±0.1	-33.1±7.5
AFC	-23.3±2.0	18.0±2.8	4.0±1.1	-24.7±5.5	-2.1±0.1	-28.0±5.6
AGG	-23.8±2.2	19.3±4.8	7.0±1.0	-35.2±7.0	-1.8±0.1	-34.5±5.6
AFG	-23.3±2.1	26.6±4.2	3.9±1.1	-30.6±5.9	-2.0±0.1	-25.4±5.4
TGA	-23.0±2.3	15.2±4.8	6.8±1.3	-28.0±7.6	-1.9±0.1	-30.9±7.2
TFA	-23.6±2.2	23.7±3.2	3.8±1.4	-29.4±6.8	-2.2±0.1	-25.8±6.9
TGC	-24.4±2.6	17.1±4.3	6.6±1.3	-33.3±6.5	-1.9±0.1	-35.9±6.2
TFC	-24.1±2.0	21.5±3.6	4.2±1.3	-28.2±6.6	-2.3±0.1	-28.9±5.4
TGT	-24.9±2.1	15.0±4.0	7.1±1.0	-30.7±7.4	-1.9±0.1	-35.4±6.2
TFT	-25.2±2.1	23.8±3.9	4.4±1.2	-26.2±7.4	-2.2±0.1	-25.4±6.4

2.4.5.2 Hydrogen bonding and stacking interactions

According to the above hypothesis, non-bonded interactions of both inter-strand (hydrogen bonding) and intra-strand (stacking interaction) of the FapydG:C or G:C base pair in different contexts of central oligonucleotides were calculated using *anal* module [109]. The resulting interactions are shown in table 2.7. The most noticeable effect of interactions is in the way that FapydG pairs with cytosine at $\approx 8\ kcal/mol$ higher than normal G:C base pairs. Although the FapydG:C base pair exhibits a G:C-like hydrogen-bonding pattern, yet the high mobility of the glycosidic bond results in the instability of those hydrogen bonds. The stacking energies of FapydG or guanine flanked by various 5'- and 3'-neighbouring bases resulted unexpectedly that the nonplanar FapydG has the tendency to generate more stable base stacking in *B*-DNA than normal guanine. Such preferences occur when the 3'-nucleobase of FapydG is adenine, cytosine or guanine whereas FapydG displays unfavourable interactions to 3'-thymine. The nonplanar formyl group can generate hydrogen bonding with an amino group in adenine and cytosine where the negative charge of formamido group may locate in a good distance and produce favourable repulsive-attractive interactions with negatively charged N7 and O6 of guanine base (see figure 2.9). The unfavourable interactions between the FapydG and 3'-thymine base step could be explained by a steric effect between the out-of-plane formamide group of FapydG and a methyl group at position 5 of 3'-thymine.

2.4.6 Intrinsic DNA curvature

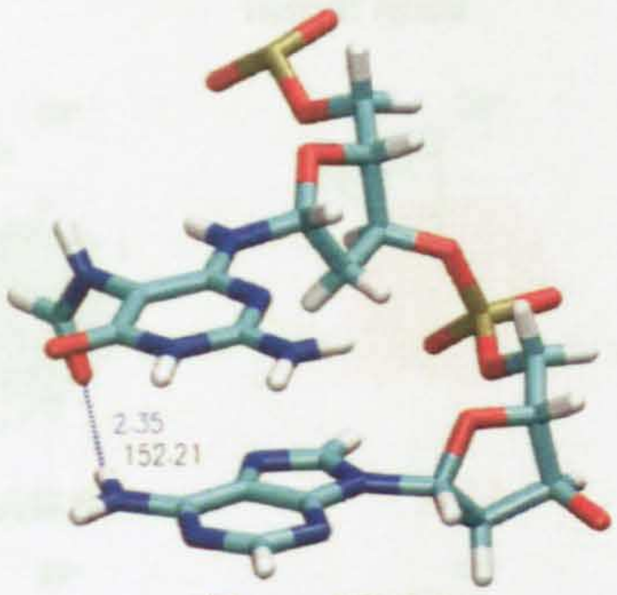
Global bending analysis provides bending information in terms of angular direction and magnitude. The sine function of DNA curvature magnitude was first plotted as a function of their angular direction associated with the DNA curvature of the kinked crystal structure (1XC8) and the short restrained simulation of the target structure illustrated by polar plots in figure 2.10. The kinked DNA conformation obtained from the complex between *Ll*Fpg and FapydG-containing

Table 2.7: Inter-strand (hydrogen bonding) and intra-strand (stacking) interactions of FapydG:C or G:C base pairs in 6 different sequence contexts.

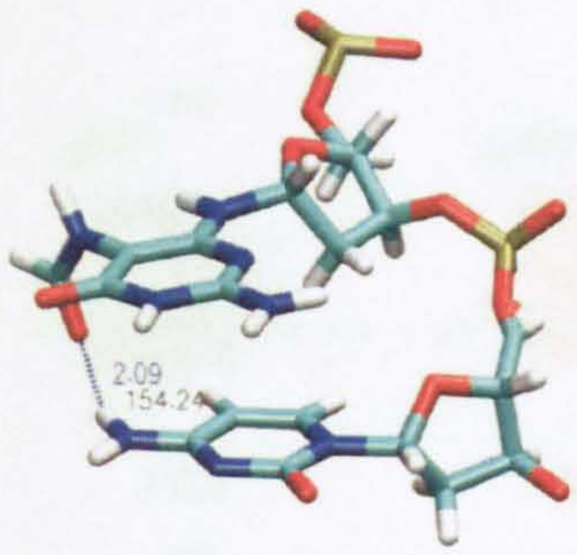
Sequence	Hydrogen bond energy (kcal/mol)	Stacking energy (kcal/mol)
AGA	-25.9±1.8	-17.6±1.7
AFA	-18.1±1.6	-19.1±1.9
AGC	-26.2±1.6	-18.8±1.8
AFC	-18.4±1.4	-23.7±1.7
AGG	-26.2±1.7	-9.7±2.0
AFG	-18.2±1.6	-10.4±1.7
TGA	-26.0±1.7	-13.7±2.0
TFA	-18.0±1.5	-13.4±1.8
TGC	-26.4±1.5	-14.7±2.0
TFC	-18.1±1.6	-17.5±1.7
TGT	-25.9±1.8	-9.9±2.0
TFT	-17.5±2.4	-8.7±2.2

DNA has indicated that Fpg is bound to the damaged DNA at an angle of curvature 63.5° and in the direction of 104.0° and this conformation has been used as a reference point in the polar plots. The restrained simulation of the kinked DNA was performed to gain some of its possible conformations required for binding to Fpg. The resulting angular magnitude and direction of the target conformations range from 51.6° to 94.7° and 103.1° to 129.3° respectively. However, these large magnitude and direction of bending rarely occurs in the absence of protein binding, though intrinsic DNA bending of only 15° to 24° is normally observed in crystal structures of *B*-DNA [84].

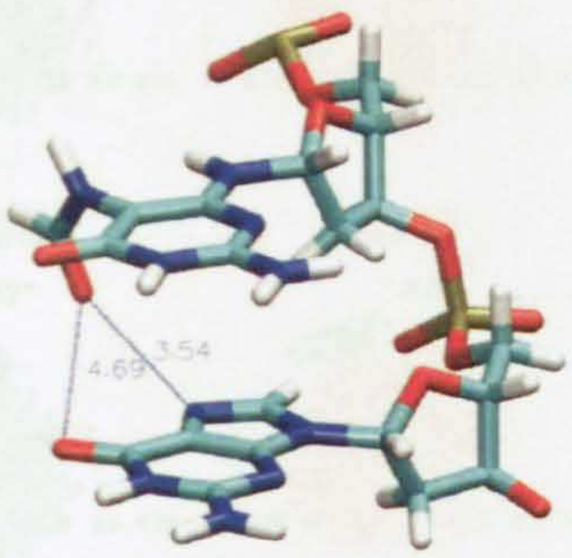
As shown in figure 2.10, the polar plots of FapydG-containing duplexes compared to its normal counterpart demonstrate the angular magnitude and direction of damaged DNA bending shifting towards the target area. It can be undoubtedly seen in the altered behaviour of AFC, AFG and TFC; AFA, TFA and particularly TFT bendability is however questionable. One-way analysis of variance (ANOVA) was used to test differences between the angular magnitude over 5-ns simulations of damaged DNA compared to undamaged DNA (α=0.05 for a 95% confidence). The results showed statistically significant differences in the



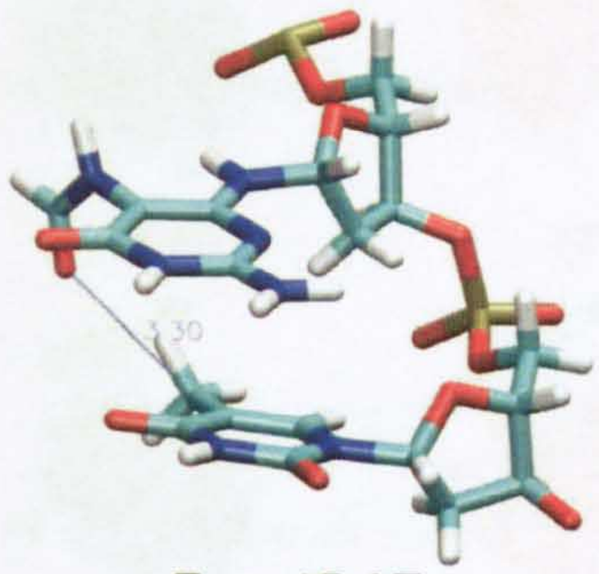
FapydG / A



FapydG / C



FapydG / G



FapydG / T

Figure 2.9: Proposed interactions between a formyl functional group of FapydG and 3'-neighbouring nucleobases (pictures generated from their time-averaged structures). FapydG/A and FapydG/C base steps show hydrogen bond lengths at 2.35 and 2.09 Å, and the angles at 152 and 154°, respectively. FapydG/G represents favourable repulsive-attractive interactions while FapydG/T indicates unfavourable steric interactions between a formamide group and a methyl group of thymine.

presence of FapydG within AXA, AXC, AXG, TXA and TXC sequences, while there was no significant difference between TGT and TFT. A histogram of DNA curvature in the presence and absence of FapydG are also shown in figure 2.11. In the presence of FapydG, shifts towards larger curvature were evident in the distribution of bending magnitude of AFA, AFC, AFG, TFA and TFC central sequences.

The average values of the angular magnitude and direction of bending of those of the damaged and the undamaged sequences over 5-ns simulations were calculated and shown in table 2.8. The average global curvature of AFA, AFC, AFG, TFA and TFC sequences demonstrated a higher value of the angular magnitude

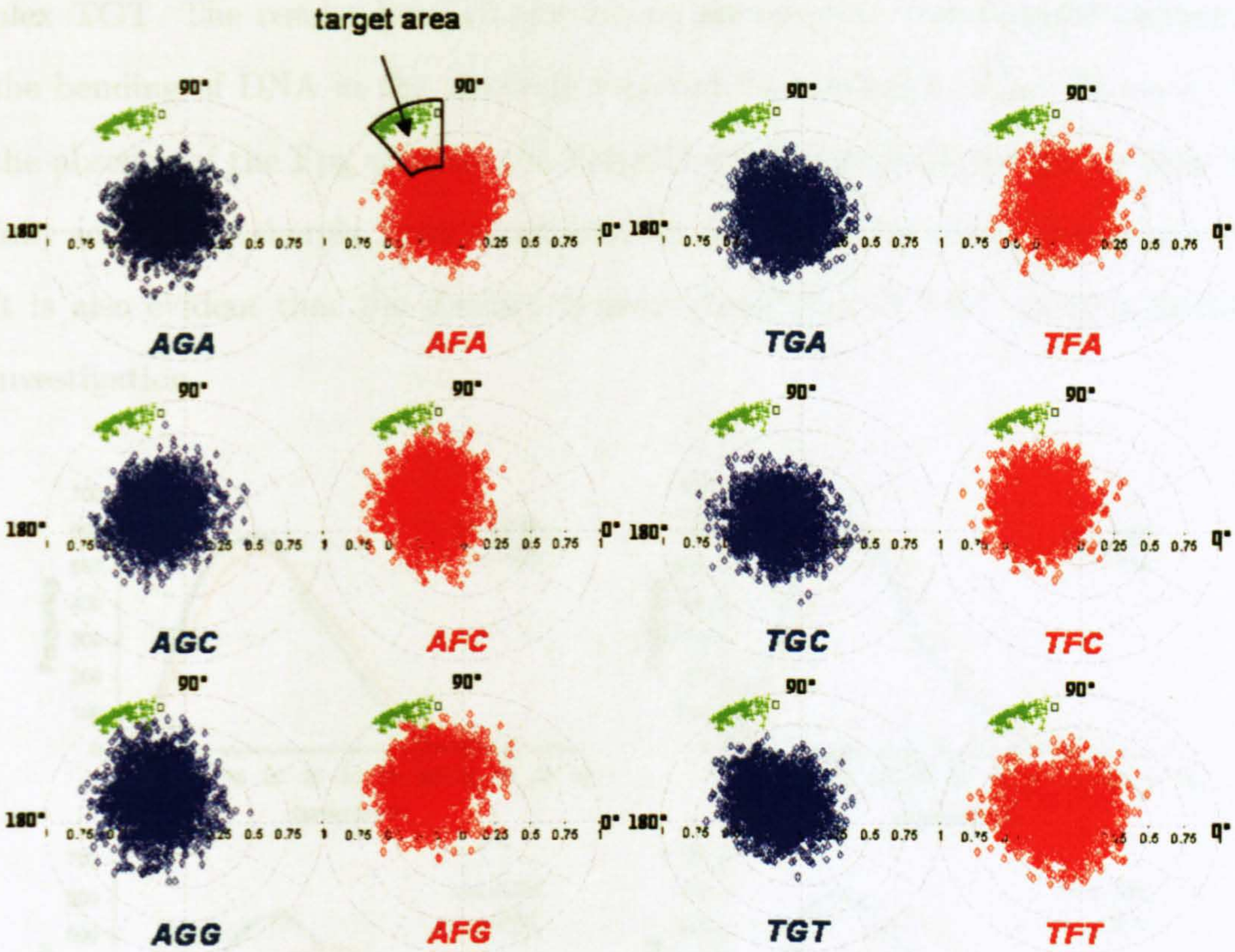


Figure 2.10: Polar plots of angular curvature ($\sin\theta$) versus direction of bending (°) of the damaged DNA (red) compared to its normal counterpart (blue) show the magnitude and direction of damaged DNA bending shifting towards the crystal structure (black square) and the target conformations (green) excluding TFT.

associated with a lower value of direction relative to its normal counterpart, whilst the TFT sequence showed the contrary dynamic behaviour. It may imply that those five damaged sequences are more likely to generate DNA conformations of the type required for binding to Fpg as in the crystal structure (PDB entry 1XC8).

To quantify the effects of FapydG on DNA flexibility, the targeted bending area was defined as the angular curvature at greater than 30° and the direction between 100° to 130°, based on the results from the restrained bent DNA simulation. It is clear that in five out of six studied sequences the proportion of conformations lying in the target area is increased when FapydG is present. Interestingly, only TFT has a lower proportion in the area than its normal du-

plex TGT. The results here all provide an assumption that FapydG enhances the bending of DNA in the direction required for binding to Fpg. However, in the absence of the Fpg protein, the FapydG-containing duplexes are unlikely to fully achieve the sharply kinked conformation as observed in the crystal structure. It is also evident that the distinct dynamic behaviour of TFT requires further investigation.

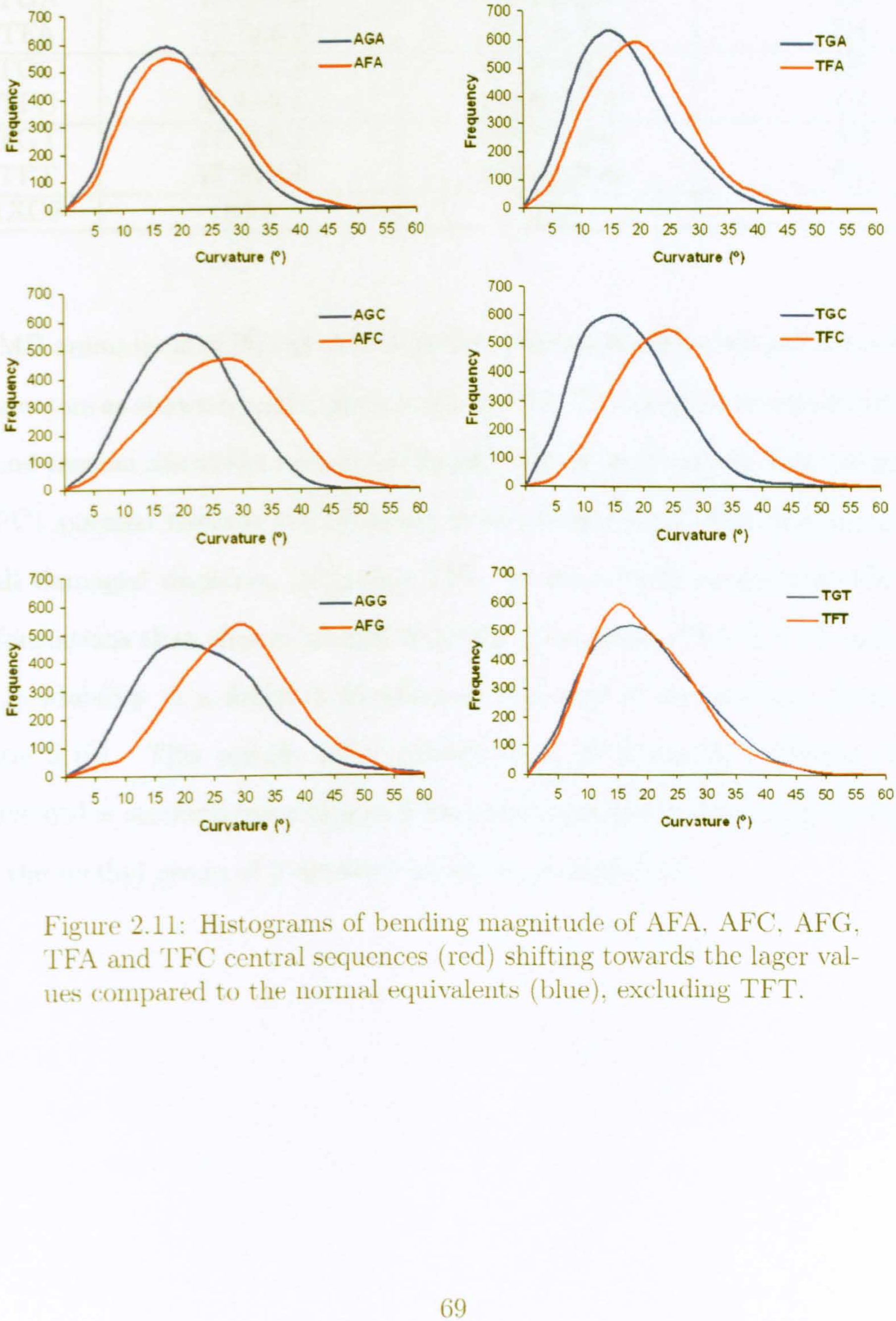


Figure 2.11: Histograms of bending magnitude of AFA, AFC, AFG, TFA and TFC central sequences (red) shifting towards the lager values compared to the normal equivalents (blue), excluding TFT.

Table 2.8: Average values of the angular magnitude and direction of global bending and proportion of the MD ensembles that show the required bending (curvature > 30°, direction = 100° - 130°) from 5-ns simulations.

Sequence	Average curvature (°)	Average direction (°)	Required bending (%)
AGA	16.2±7.8	135.6±56.8	1.3
AFA	18.1±8.7	126.2±48.7	4.0
AGC	18.4±8.2	141.7±48.7	2.1
AFC	23.7±9.8	138.2±38.4	12.2
AGG	20.9±10.1	137.6±48.0	7.1
AFG	26.0±9.5	128.0±32.8	12.4
TGA	15.1±7.8	137.6±70.2	2.0
TFA	17.7±8.2	116.7±46.5	3.4
TGC	15.0±7.9	158.6±65.5	0.6
TFC	22.3±9.1	146.0±32.8	4.4
TGT	17.4±9.1	136.8±55.4	3.2
TFT	17.0±8.7	151.5±66.6	0.7
1XC8	63.5	104.0	-

MD animations of PC1 of each sequence were analysed for the global curvature parameters as shown by polar plots in figure 2.12. The analysis revealed that PC1, a bend motion about the lesion site, mainly contributed towards Fpg recognition as PC1 pointed towards the direction of the target area. The first component of all damaged duplexes, excluding TFT, is more likely to generate the bent conformations than that of normal duplexes. Noticeably, PC1 of TFT influences the bendability in a different direction as observed in the previous polar plots (figure 2.10). This specific DNA behaviour of TFT can be explained by the unfavourable stacking interactions of the nonplanar formamide group of FapydG and the methyl group of 3'-thymine as shown in figure 2.9.

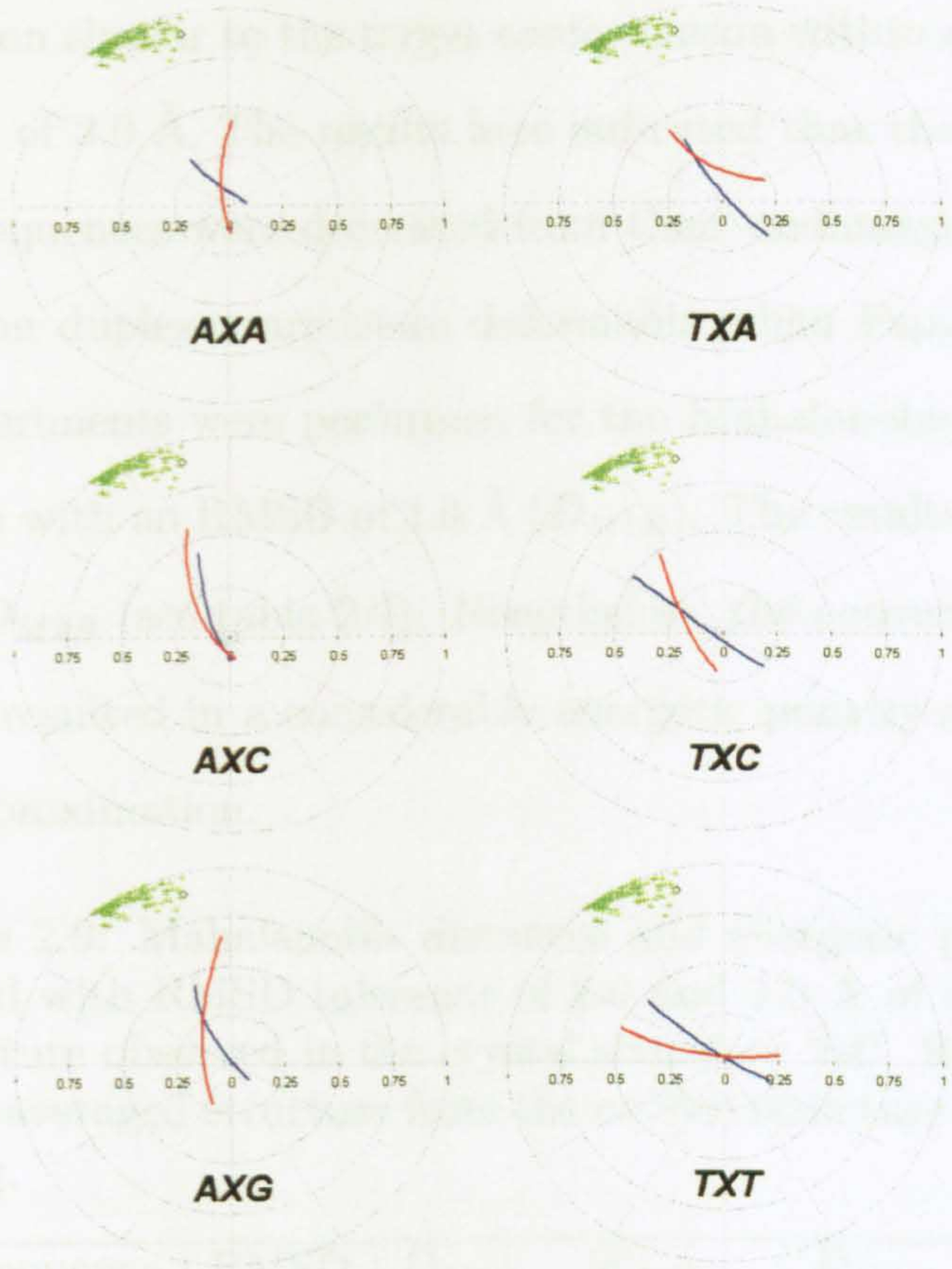


Figure 2.12: Polar plots of angular curvature (sin θ) versus direction of bending (°) of the major mode of DNA deformation (PC1) of the damaged DNA (red) compared to their normal counterparts (blue).

2.4.7 Mahalanobis distances

The D_M was used to determine the deformability towards the protein-bound conformation. It is not only measure the similarity between two configurations of a system as in the conventional RMSD method but it also takes into account of major modes of deformation of DNA and roughly reflects the energy (E_{def}) required to deform the structure from the average configuration towards the target using a Hooke model. In general, the greater the value of D_M, the higher energy is required to deform towards the target configuration. The results of D_M and E_{def} are demonstrated in table 2.9.

The $D_{M2.0}$ is the lowest distance from the time-averaged structure to adopt a conformation similar to the target conformation within a conformational space of an RMSD of 2.0 Å. The results here indicated that the $D_{M2.0}$ of all FapydG-containing sequences were decreased from their undamaged equivalents. It suggests that the duplexes are more deformable when FapydG is present. More rigorous experiments were performed for the Mahalanobis distance to achieve a conformation with an RMSD of 1.0 Å ($D_{M1.0}$). The results of $D_{M1.0}$ were similar to those of $D_{M2.0}$ (see table 2.9). Nonetheless, the conversion of $D_{M1.0}$ into the energy term resulted in a considerably energetic penalty as it was based on the harmonic approximation.

Table 2.9: Mahalanobis distances and energetic penalties associated with RMSD tolerance of 2.0 and 1.0 Å of the bent DNA structure observed in the crystal structure [54]. RMSDs of each time-averaged structure from the crystal structure are also listed along.

Sequence	RMSD (Å)	$D_{M2.0}$	$E_{def2.0}$ (kcal/mol)	$D_{M1.0}$	$E_{def1.0}$ (kcal/mol)
AGA	4.20	5.33	8.46	15.55	72.06
AFA	4.09	4.90	7.15	14.45	62.22
AGC	4.10	5.85	10.2	18.64	103.5
AFC	3.77	3.90	4.53	14.38	61.62
AGG	3.98	4.30	5.51	13.29	52.63
AFG	3.84	4.25	5.38	13.21	52.00
TGA	4.26	4.62	6.36	13.92	57.74
TFA	3.94	4.20	5.26	11.93	42.41
TGC	4.12	5.15	7.90	18.34	100.2
TFC	3.87	3.94	4.63	13.97	58.16
TGT	4.18	4.93	7.24	15.96	75.91
TFT	4.38	4.84	6.98	15.60	72.52

It is noticeable that $D_{M1.0}$ and $D_{M2.0}$ of TFT are relatively higher than other damaged sequences indicating that a higher energetic penalty is required to distort it to the protein-bound conformation. The result here has a good agreement with the previous axis curvature analyses that TFT tends to deform not in the required direction. This is in contrast to the AFC and TFC sequences, which require a lower energetic penalty to achieve the target conformation, in another word, they

are more deformable than the TFT sequence. This individually intrinsic curvature may result in differences in Fpg activity to recognise the lesion flanked to 3'-cytosine more efficiently than the lesion with 3'-adjacent thymine. These findings support the hypothesis that the damage recognition is sequence-dependent and may be related to the energetic cost of DNA distortion, as has been reported in the activity of uracil DNA glycosylase [116]. Although glycosylases are expected to recognise and remove DNA damage in any sequence context, there have been some reports that the UDG activity is dependent on the sequence surrounding the uracil with up to 20-fold differences in UDG efficiency [117, 118, 119].

Regarding to the D_M of damaged and undamaged sequences, the D_M of AFC and TFC sequences are dramatically decreased whereas those of AFG and TFT contexts are slightly dropped, compared to their normal counterparts. If the repair enzyme could detect the different flexibility in the presence and absence of FapydG, it would be able to distinguish or recognise FapydG flanked to 3'-cytosine better than in other sequence contexts. Alternatively, it may suggest that the effect of FapydG in TFT and AFG flexibility is concealed from the repair process of Fpg.

2.5 Conclusions

Parameterisation of FapydG and verification of the potential glycosidic conformation were initially resolved before carrying out MD simulations. The simulations were then performed on 12-mer oligonucleotides containing *anti*-FapydG:C and G:C in various neighbouring bases. The PCA approach was employed to extract essential information from a massive numerical output from simulations. DNA bending at the lesion site was the major mode of motion (PC1) where damaged and undamaged duplexes deformed in different extents. Energetic analysis showed the destabilisation of FapydG:C due to the weaker hydrogen bonding that may lead to base flipping into the pocket of Fpg afterwards. However, there was no tendency of base opening during our simulations.

DNA conformations were analysed in terms of the global axis curvature and direction. In the presence of FapydG, the direction and magnitude of global bending are both significantly more likely to generate DNA conformations of the type required for binding to Fpg. Nonetheless the sharply kinked duplex seems to be unachievable without the protein binding. Results of these studies provide the validation of DNA bending enhancement by oxidative base lesions, which can be observed from PC1. It is highly likely that DNA deformation is one principal element of how the repair protein distinguishes the lesion from the vast expanse of normal bases.

The distinct behaviour of TFT was firstly noticed by analysing its angular magnitude and direction showing the impediment of the flexibility. We hypothesised that the steric interaction between a nonplanar formamide group of FapydG and a methyl group at position 5 of successive 3'-thymine inhibits the DNA flexibility. This behaviour may impair the recognition process and the efficiency of repair enzymes.

Chapter 3

Damage Recognition by DNA Glycosylases

3.1 Introduction

DNA base damage is typically excised from the genome by DNA glycosylases. The glycosylase must stabilise the damaged base in an extrahelical conformation inside its lesion-specific recognition pocket forming appropriate contacts and orientation before excision can take place. This finding raises fascinating questions of whether the enzymes need to extrude each base into the recognition pocket in order to distinguish the lesion from native nucleobases, or whether the proteins are able to locate and flip only the damaged base which is situated within the DNA duplex [120]. To understand how the lesion recognition pocket discriminates between damaged and undamaged bases, it is necessary to have structural information of the protein binding pocket when bound to DNA in the presence of a lesion or its normal counterpart. Unfortunately, DNA-binding proteins commonly fail to bind specifically to DNA without cognate groups such as lesions or promoter sequences. To grant the missing knowledge, theoretical studies are required to construct models of the protein bound to non-lesion DNA based on all sources of information available to date.

3.1.1 FapydG-specific interactions with Fpg

FapydG, as well as 8-oxoguanine (8OG), is a well-known substrate for the bacterial formamidopyrimidine-DNA glycosylase (Fpg or MutM). Although other lesions such as 8-oxoadenine (8OA), 5-hydroxycytosine (5OHC) and dihydrouracil (DHU) are also substrates for Fpg, those lesions are excised less efficiently than FapydG by Fpg [54]. To investigate the recognition mechanisms of FapydG by Fpg at an atomic level, it was initially probed based on available Fpg crystal structures deposited in the Protein Data Bank (PDB) [55].

Four different species of bacterial Fpg proteins have been used for crystallographic studies: *Thermus thermophilus* (*Tt*) Fpg [121], *Escherichia coli* (*Ec*) Fpg [122], *Bacillus* (or *Geobacillus*) *stearothermophilus* (*Bst*) Fpg [123] and *Lactococcus lactis* (*Ll*) Fpg [56, 54]. Most of these crystal structures are in DNA-bound enzyme complexes except for a structure of free *Tt*Fpg protein. Unfortunately, the absence of structures of both uncomplexed and DNA-bound enzymes from the same source has hindered damage recognition studies. Another bacterial enzyme which shares significant sequence similarity with the bacterial Fpg protein is the bacterial endonuclease VIII (EndoVIII or Nei) [124]. EndoVIII efficiently repairs a number of oxidised pyrimidines, including thymine glycol (Tg), dihydrothymine (DHT) and DHU [125, 126, 127]. These two subfamilies are categorised as the same structural Fpg superfamily. The global structure of the Fpg superfamily consists of two distinct domains connected by a flexible hinge with a large, electrostatically positive cleft lined by highly conserved residues between the domains. Amino acid sequences and their alignment of the Fpg superfamily have been shown in figure 3.1.

The most remarkable crystal structure for FapydG recognition studies is the wild-type *Ll*Fpg-DNA containing *c*FapydG complex (PDB entry 1XC8, 1.95 Å resolution) [54] as previously described in 1.2. The extrahelical conformation of the *anti*-glycosidic *c*FapydG lesion was constrained inside the recognition binding pocket of Fpg. The pocket was formed between N- and C-terminal domains of

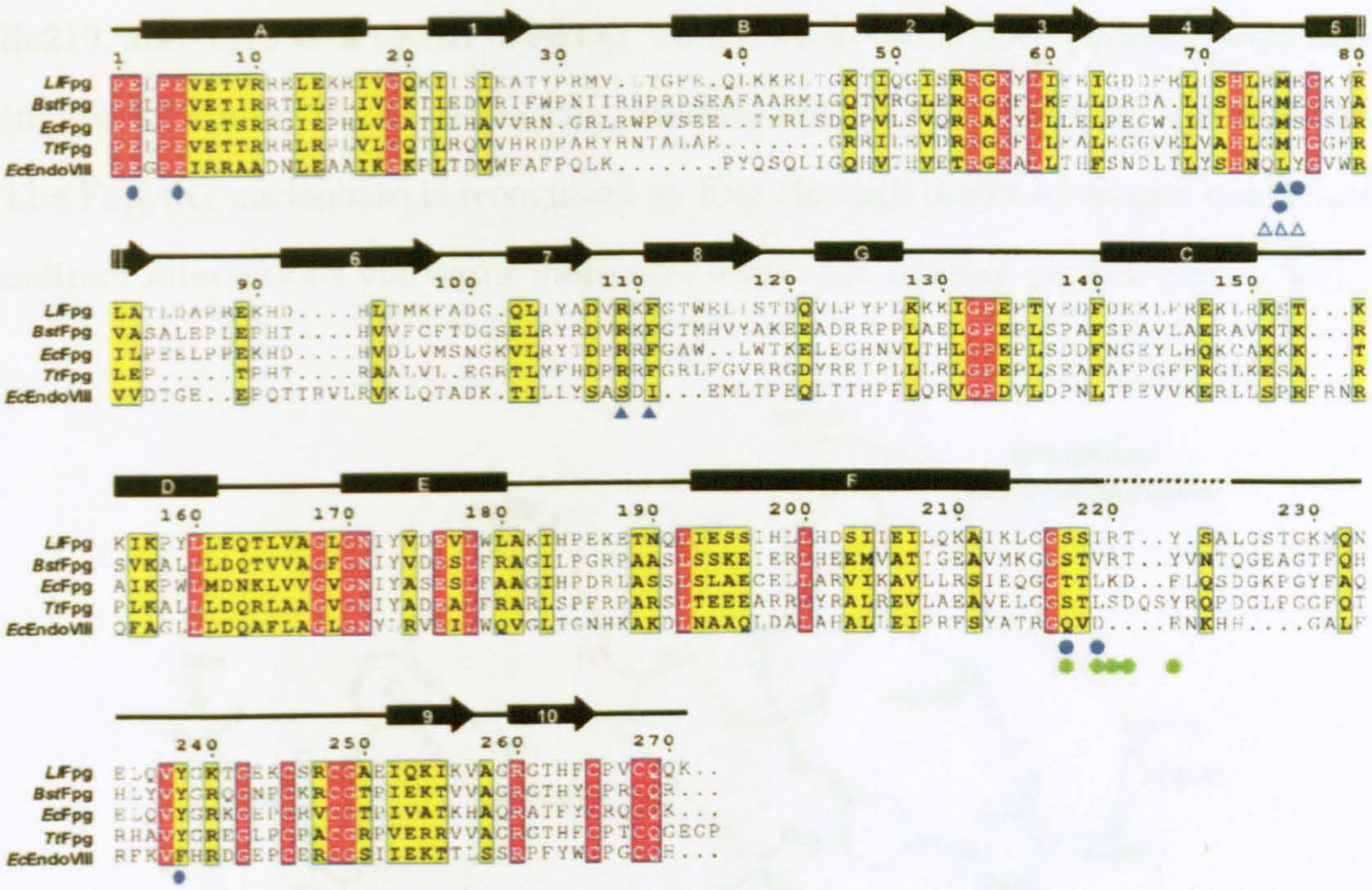


Figure 3.1: Primary sequence alignments of the Fpg superfamily. The *bold black* characters in *yellow boxes* indicate the homologies. The *closed* and *open triangles* represent the residues that are intercalated in DNA at the target site for Fpg and EndoVIII, respectively. The residues of the binding pocket involved in the recognition of FapydG and 8OG are indicated by *blue* and *green* circles, respectively. The β-strands and α-helices are illustrated by *black arrows* and *rectangles*, respectively. The *black dashed line* represents the missing residues in the electron density map. The figure is taken from [54].

the enzyme by the α-helix A and the β4-β5 loop of the N-terminus and the αF-β9 loop located between the H2TH motif and the zinc finger in the C-terminus. Interestingly, the flexible part of the αF-β9 loop (residues 220-224), which was disordered in the first published X-ray model of *Ll*Fpg bound to a *c*FapydG-containing DNA (PDB entry 1TDZ, 1.8 Å resolution), was suggested to play a major role in the lesion recognition by creating a hydrogen bonding network between either the main chain or the side chain of the protein and part of FapydG [54, 128].

Based on the first publication of the *Ll*Fpg/DNA complex [54], several Fpg residues appear to be involved in the specific recognition of the FapydG lesion: Glu2 and Glu5 of the α-helix A, Met75 and Glu76 of the β4-β5 loop, and Ser217,

Ile219, and Tyr238 of the αF-β9 loop. As shown in the aligned primary sequences (figure 3.1), most of the interacting residues are strictly conserved except Glu5. The FapydG nucleobase is recognised by Fpg through direct hydrogen bonds and indirect interactions via water molecules inside the binding pocket (figure 3.2).

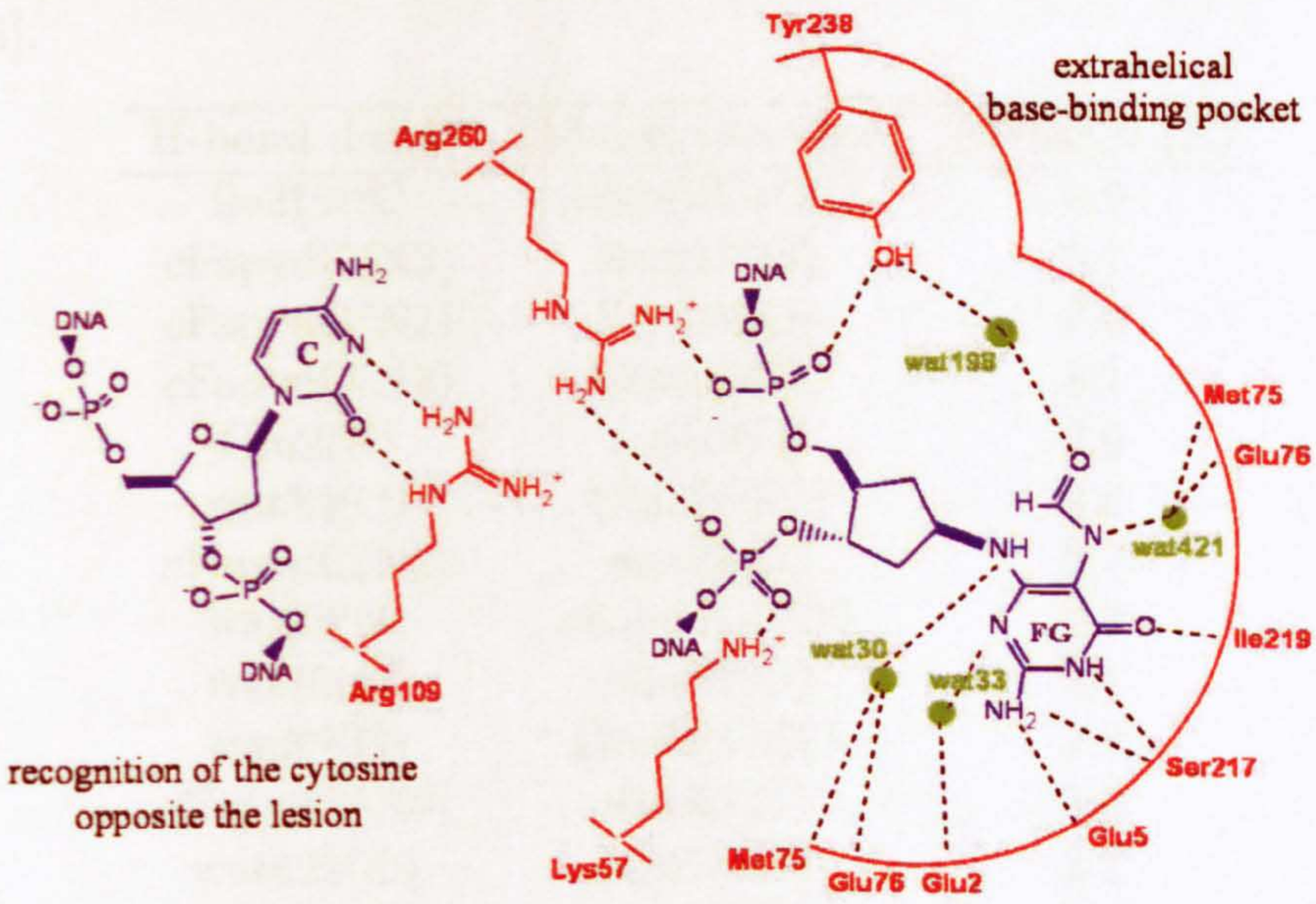


Figure 3.2: A schematic diagram of Fpg/DNA contacts at the binding site. Amino acid residues, nucleotides and waters involved in the recognition are shown in *red*, *blue* and *green*, respectively. The figure is reproduced from [54].

The recognition of the N1 atom of FapydG is through hydrogen bonds with the main chain amide and the side chain carboxyl groups of Glu2 through wat33. The N2 amino group of FapydG is directly contacted by either the side chain carboxyl group of Glu5 or the main chain carbonyl group of Ser217, while the main chain of Ser217 also forms a hydrogen bond with N3 of FapydG. The O4-carbonyl of FapydG is recognised by the main chain amide of Ile219. The formamide functional group of FapydG is indirectly hydrogen bonded by the main chain of Met75 and the side chains of Glu76 and Tyr238 through wat421 and wat198. Met75 and Glu76 are also involved in the recognition of the glycosidic amino N6 group of FapydG through water molecule-assisted interactions. Three water

molecules (wat30, wat33 and wat421) also present within the binding pocket of previous crystal structures of Fpg-DNA complexes [122, 129]. Table 3.1 summarises distances between hydrogen bond donors and acceptors of *c*FapydG and the surrounding Fpg residues inside the binding pocket.

Table 3.1: Distances between hydrogen bond donors and acceptors of Fpg/DNA contacts at the binding site. The table is adapted from [54].

H-bond donor*	H-bond acceptor*	Distance (Å)
Ile219(N)	*c*FapydG(O4)	2.9
*c*FapydG(N3)	Ser217(O)	2.7
*c*FapydG(N2)	Ser217(O)	3.0
*c*FapydG(N2)	Glu5(OE2)	3.1
Glu2(N)	wat33(O)	3.0
wat33(O)	Glu2(OE2)	2.8
*c*FapydG(N2)	wat33(O)	2.7
wat33(O)	*c*FapydG(N1)	2.7
wat30(O)	Met75(O)	3.1
wat30(O)	Glu76(OE2)	2.7
*c*FapydG(N6)	wat30(O)	3.0
wat421(O)	Met75(O)	3.2
Glu76(OE1)	wat421(O)	3.3
*c*FapydG(N5)	wat421(O)	2.5
Tyr238(OH)	wat198(O)	2.9
wat198(O)	*c*FapydG(O)	2.6

*Atom names of hydrogen donor and acceptor residues are indicated in parentheses.

In addition to the specific interactions to constrain *c*FapydG in the recognition binding pocket, three residues in the N-terminal domain (Met75, Arg109 and Phe111) are intercalated into the DNA backbone via the minor groove, preventing the lesion reinsertion into the DNA duplex. The intrahelical cytosine opposite cFapydG is also stabilised by the formation of hydrogen bonds with Arg109. The DNA phosphodiester backbone of the damaged strand is also maintained by the cooperation of positively charged residues establishing the inner surface of the pocket for DNA binding. The sharply kinked DNA is constrained at the lesion site by Lys57, Tyr238 and Arg260, which collectively contact the two phosphate groups adjacent to the site.

3.1.2 Aims and objectives

After being oxidised by reactive oxygen species (ROS), the imidazole ring of guanine is ruptured resulting in the formation of the formamide functional group and the increased degree of freedom in the glycosidic bond. Undoubtedly, these functional groups are explicitly different from guanine and are of interest in this study. The aims of the work in this chapter were attempted to improve the proposed Fpg-FapydG interaction model from the previous report [54] and to study how the recognition pocket distinguishes the FapydG lesion from its normal counterpart through these major chemical alterations. The dynamics of the αF-β9 loop that was missing in the electron density map of the crystal complex was also a particular region of interest since this flexible loop may establish specific interactions with the formamide group of FapydG.

MD simulations of the wild-type *Ll*Fpg-DNA containing *c*FapydG complex and its undamaged model with an extrahelical guanine were performed. The undamaged model was constructed by the replacement of FapydG by its normal equivalent at the same location in the binding pocket. Specific interactions between the distinct functional groups of the FapydG lesion and Fpg residues inside the recognition pocket were first investigated during the simulations compared to the non-lesion complex. The influence of FapydG on protein flexibility of the αF-β9 loop compared to its normal equivalent was also studied. Relative binding free energies of the damaged and undamaged complex were finally calculated using the MM/GBSA approach.

3.2 System preparation and simulation

3.2.1 Protein system setup

Coordinates of the X-ray crystallographic structure of the Fpg protein bound to a FapydG-containing oligonucleotide (Fpg/FG complex) and 4 structural water molecules inside the binding pocket (wat30, wat33, wat198, and wat421 in

figure 3.2) were initially obtained from PDB entry 1XC8 [54]. The sequence of the DNA duplex was d(TCTTTFTTTCTC)·d(GAGAAACAAAGA), where F is FapydG. The Fpg model was thoroughly prepared prior to performing simulations. Protonation states of Fpg residues were first decided using the web interface version of the WHAT IF suite of program [130]. The result suggested reassignments of histidine residues as follows: His201 was renamed as protonated histidine (HIP), His198 and His263 as δ-histidine (HID), while all others were assigned as ϵ-histidine (HIE). The modified Fpg model showed the improvement of its stability from the original structure through an increasing number of the intramolecular hydrogen bonds. Furthermore, the N-terminal proline residue was modified as neutral to mimic the pre-catalytic stage before base excision since it was suggested that the glycosylase activity of the DNA glycosylase family is initiated by attacking of the nucleophilic residue on the C1'-deoxyribose moiety of the target [131, 132]. Force field parameters for the neutral N-terminal proline were obtained from [114].

Finally, the zinc finger motif was disordered during MD simulations when the original nonbonded zinc parameters ($\sigma = 1.10$ Å, $\epsilon = 0.0125$ *kcal/mol*) from the AMBER force field were used. Thus, a longer-range electrostatic model ($\sigma = 1.70$ Å, $\epsilon = 0.67$ *kcal/mol*) of zinc ion (Zn^{2+}) as reported by Stote and Karplus was employed [133]. The Zn^{2+} was coordinated within the zinc finger motif by four cysteine residues 245, 248, 265 and 268, at those were renamed as CYM (cysteine with negative charge). Geometry of tetrahedrally coordinated Zn^{2+} in the centre of the zinc finger motif was subsequently maintained throughout the simulations. The remaining protein and nucleic acid parameters were determined using the AMBER ff03 force field [66].

3.2.2 Undamaged model construction

Regarding the construction of the non-lesion complex model (Fpg/G complex), it has been recently documented that the extrahelical guanine is trapped outside

the active site of the repair enzyme hOGG1 using a disulphide-crosslinking strategy [134]. The crystal structure (PDB entry 1YQK) showed that DNA backbone of the undamaged system after being crosslinked was drastically bent, even more than its damaged system (hOGG1-8OG containing DNA complex, PDB entry 1EBM). However, the biological relevance of this study may be questioned. The disulphide crosslink between the estranged cytosine and hOGG1, in fact, may result in the anomalous backbone conformation when bound to the enzyme. Thus, in this study, the extrahelical FapydG conformation was replaced by guanine at the same location using the *Leap* module regardless of the binding of guanine at the alternative site. The initial structure of the extrahelical guanine was nicely superimposed on FapydG inside the binding pocket and was stabilised by a hydrogen bonding network similar to that of the FapydG residue.

3.2.3 Simulation conditions

Both Fpg/FG and Fpg/G complexes were solvated with the explicit TIP3P model in truncated octahedral boxes with a minimum of 8 Å buffer between the box edge and any solute atom, while the water molecules inside the binding pocket from the crystal structure were retained. Minimisation, equilibration and 10-ns MD production were performed as previously described in section 1.4.2 associated with 5 $kcal/mol/Å^2$ cartesian restraint on both terminal base pairs in order to maintain the distorted DNA conformation.

3.3 Post simulation analysis

3.3.1 Structural analysis

The RMSD time series with respect to the initial conformation of three regions of the damaged and undamaged complexes were separately analysed: (i) protein, which includes the peptide backbone atoms (N, C_α, C and O) of all protein residues; (ii) the lesion site, which is calculated using all heavy atoms of only

the FapydG:C or G:C and flanking base pairs; and (iii) the αF-β9 loop, which includes the peptide backbone atoms of residues in the base recognition loop (216-224). The relative atomic fluctuations around the average position of the peptide backbone atoms of all protein residues and all heavy atoms of the helix were calculated by RMSF. All these analyses were carried out using the *ptraj* module.

3.3.2 Relative binding energies

The MM/GBSA method, as previously described in section 2.3.2.1, was employed for calculation of relative free energies of binding between the Fpg binding pocket and the extrahelical base. In this chapter, the protein and DNA (without FapydG or G) was defined as the "receptor" and the FapydG or guanine nucleobase as the "ligand". The GB^{OBC} solvent model with *igb*=2 was used as suggested by by Kormos and Beveridge for protein simulations [107].

3.4 Results and discussions

3.4.1 Structural and energetic analysis

Simulations were performed for 10 ns for each of the damaged and undamaged systems. RMSD time-series throughout each trajectory was performed for three components; the protein backbone, the lesion site and the αF-β9 loop, as shown in figure 3.3.

The RMSDs of all components fluctuate in the region under 1.7 Å suggesting that both Fpg/FG and Fpg/G complexes are stable during the simulations. The protein backbone in the Fpg/FG system is likely to be slightly more flexible than that of the Fpg/G complex with an averaged RMSD at 1.1 and 1.0 Å, respectively, whereas the αF-β9 loop of the Fpg/G complex is more flexible than that of Fpg/FG with an averaged RMSD at 0.8 and 0.6 Å, respectively. One possible explanation for these observations would be that the increased flexibility

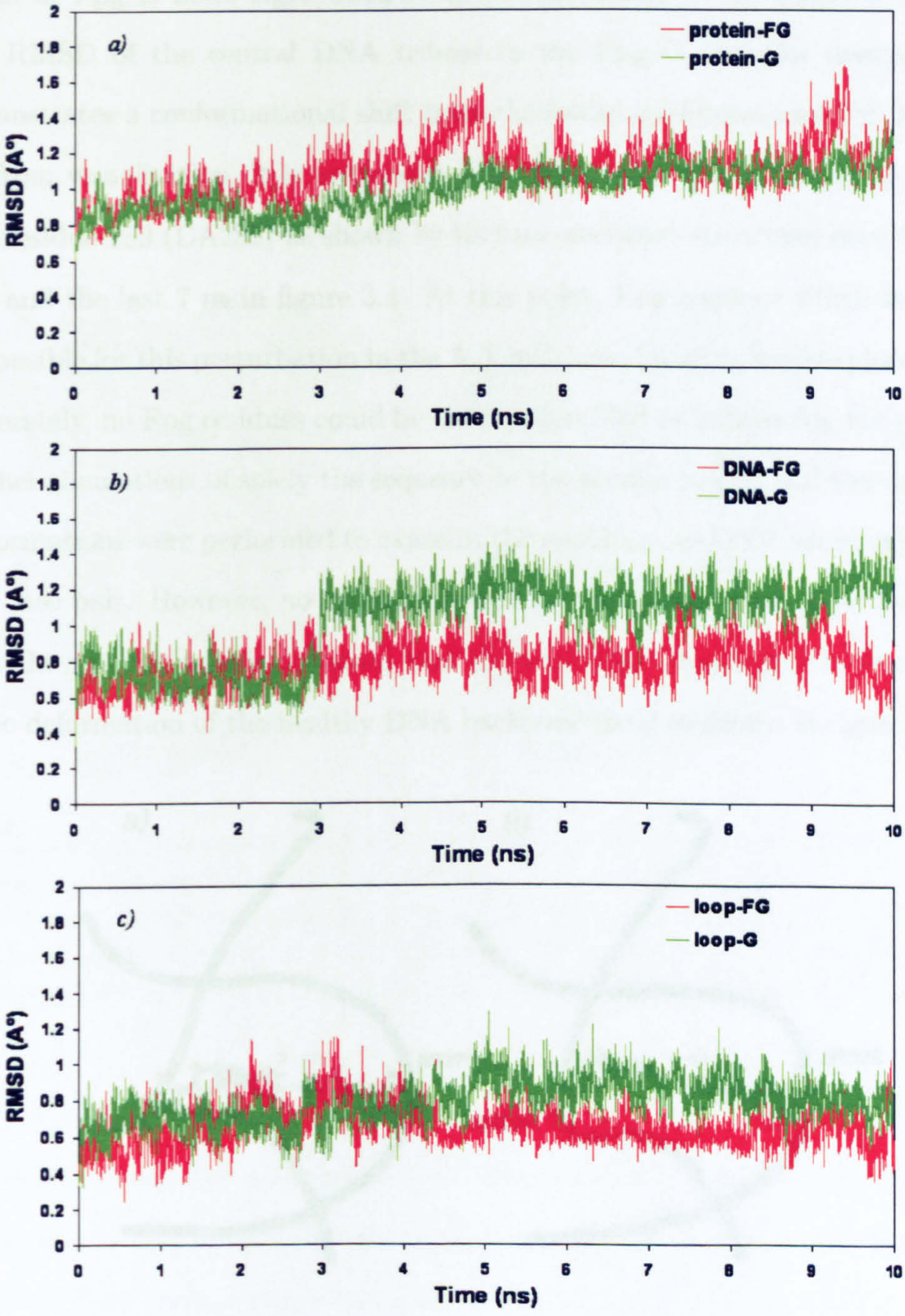


Figure 3.3: RMSD plots in Å for (a) the protein, (b) the DNA tribase, and (c) the αF-β9 loop compared to the starting structure during the simulations of Fpg/FG (*red*) and Fpg/G (*green*) complexes.

of the whole Fpg structure complexed with the damaged DNA is to permit it to rearrange itself for accommodation of its specific substrate DNA while the loop region of Fpg is more rigid when it specifically binds to the FapydG residue. The RMSD of the central DNA tribase in the Fpg/G complex unexpectedly demonstrates a conformational shift from the initial conformation after 3 ns.

From visualisation, it is noticeable that there is base pair breathing of adenine residue 289 (DA289) as shown by its time-averaged structures over the first 3 ns and the last 7 ns in figure 3.4. At this point, Fpg residues which might be responsible for this perturbation in the A:T hydrogen bonding were explored. Unfortunately, no Fpg residues could be clearly identified as influencing the process. Further simulations of solely the sequence in the normal helical and the distorted conformations were performed to examine the spontaneous DNA breathing at the A:T base pair. However, no evidence of the breathing was seen in both simulations. In conclusion, it is proposed that the breathing event is due to fine details of the deformation of the healthy DNA backbone itself as shown in figure 3.4(b).

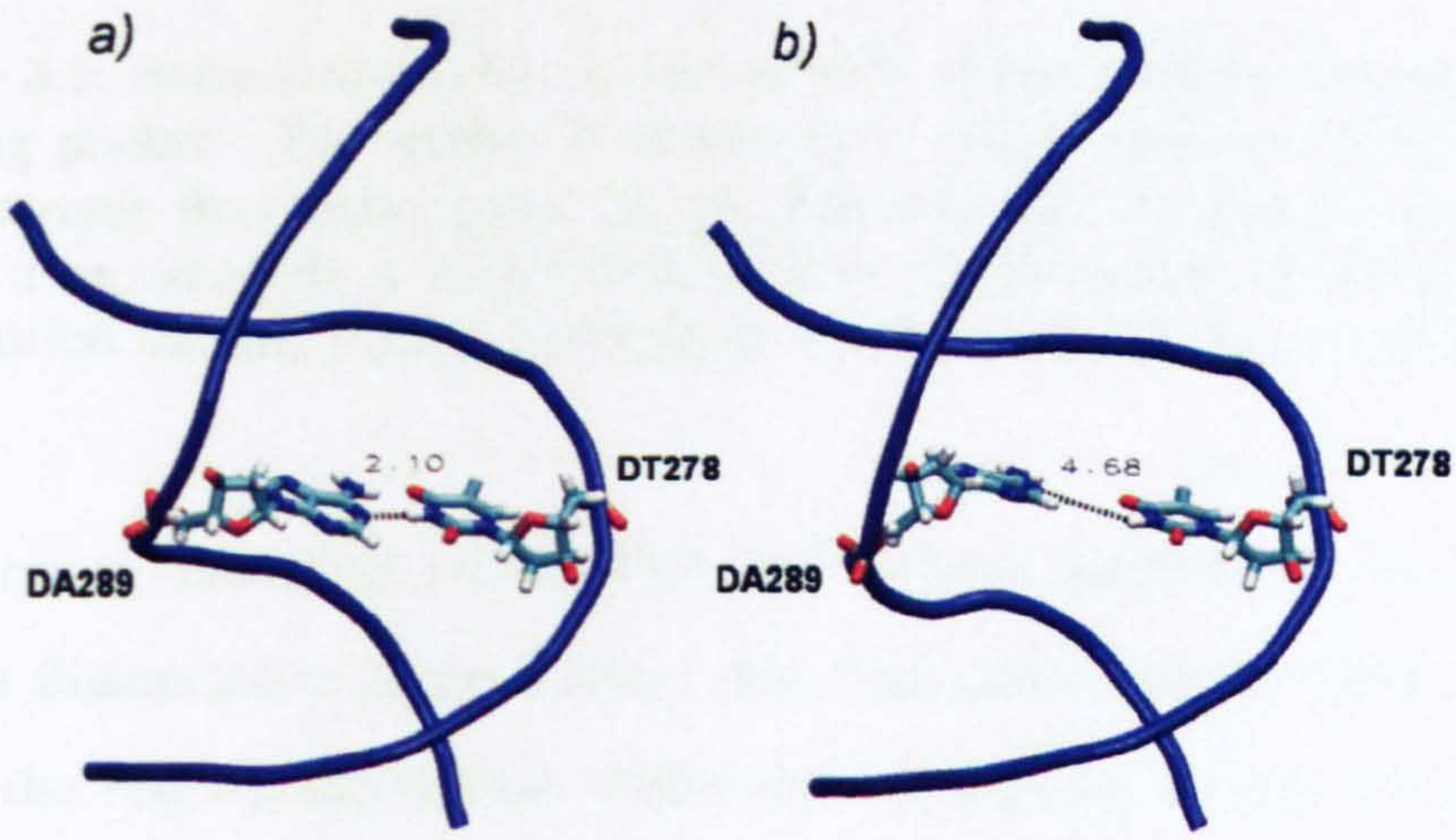


Figure 3.4: Comparisons of time-averaged structures of (a) the first 3 ns and (b) the last 7 ns of DNA conformation from Fpg/G complex showing spontaneous breathing of the DA289 residue on the healthy DNA backbone.

The relative atomic fluctuations of the protein and DNA backbone were subsequently calculated by RMSF to investigate the molecular flexibility over the simulation time. The protein flexibility was first investigated and is depicted in figure 3.5. Colours range from red where the residues are highly rigid, to green when the highly flexible residues occur. It can be seen clearly that the binding pocket in Fpg/FG complex is more rigid than that of the non-lesion complex, particularly in the αF-β9 loop.

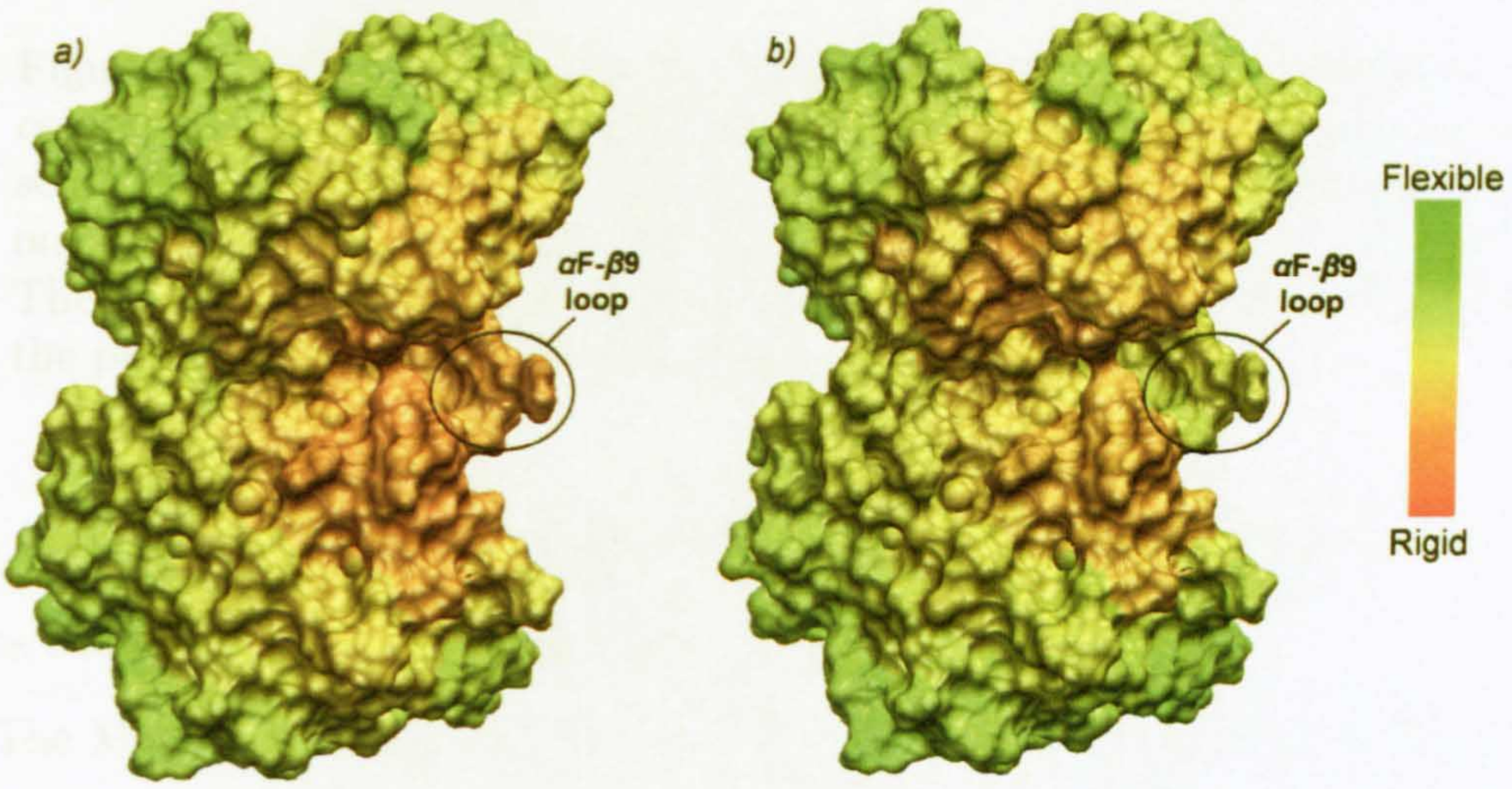


Figure 3.5: Solvent-accessible surface models of Fpg looking towards the binding pocket. The surface is shown by a colour gradient of the relative atomic fluctuation from the (a) Fpg/FG and (b) Fpg/G simulations, demonstrating a major difference in the flexibility of the lesion-recognition binding pocket particularly the dynamics of the αF-β9 loop.

Furthermore, flexibility of the DNA and enzyme backbone in the Fpg/FG complex is illustrated in figure 3.6(a). The FapydG-containing DNA is firmly bound to the Fpg binding pocket whilst there is a larger amount of movement of the guanine residue inside the binding pocket figure 3.6(b). It suggests that the FapydG residue is preferred to guanine inside the active site pocket - in other words, FapydG is accommodated by the binding pocket better than its equivalent guanine. The dynamics of the αF-β9 loop is likely to play an important role to form the precise geometry for the lesion binding pocket. The unfavourable interactions between the flipped-out guanine and the pocket in the Fpg/G complex

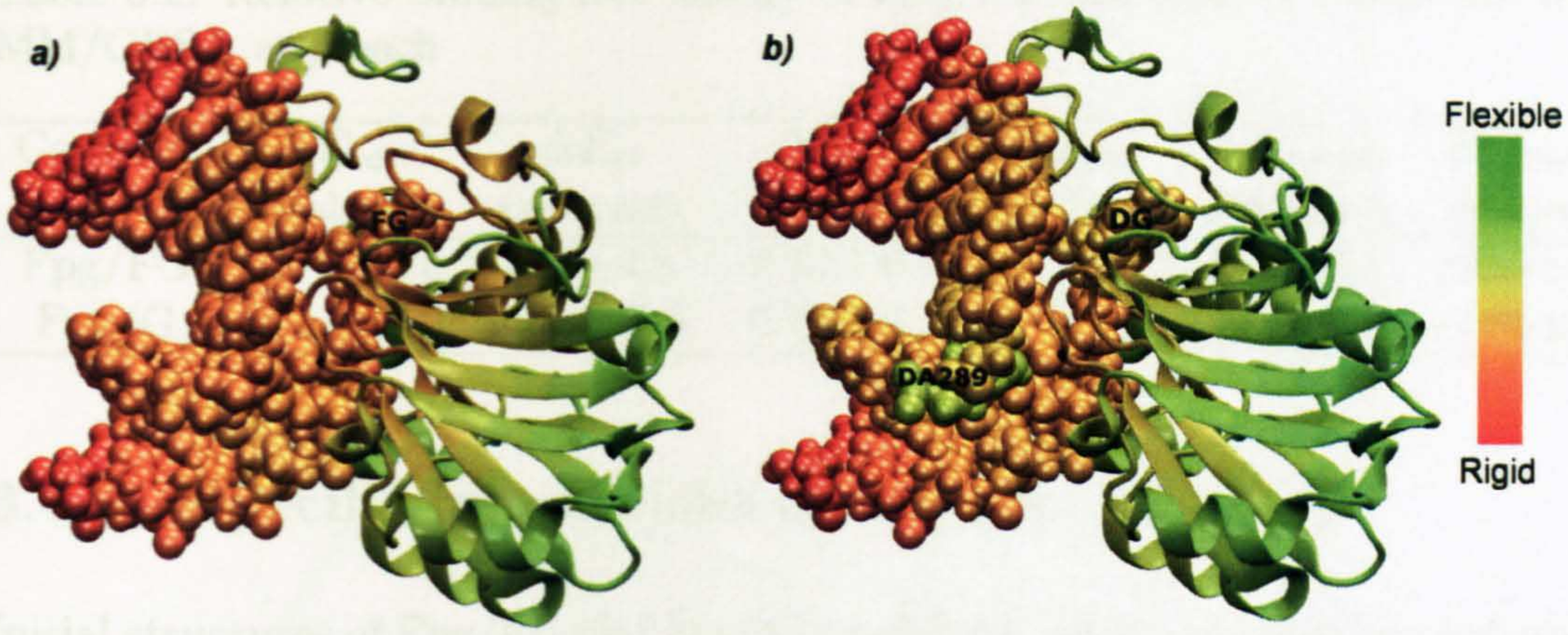


Figure 3.6: Comparisons of (a) Fpg/FG and (b) Fpg/G complexes coloured by the relative atomic fluctuation from their simulations represents the more flexibility of the Fpg/G complex, particularly the flipped-out guanine and the DA289 residue, compared to the Fpg/FG complex. The DNA conformation is represented by the van der Waals model, and the protein is with a cartoon diagram.

may additionally lead to the higher mobility of the kinked DNA structure as well as its detectable base breathing of the DA289 residue.

The MM/GBSA approach was used to further calculate relative binding free energies between the enzyme and the extrahelical FapydG compared to its normal counterpart (table 3.2). Although the electrostatic interaction (ΔE_{es}) seems to have a major contribution to stabilising guanine inside the pocket, it is neutralised by the van der Waals interaction (ΔE_{vdW}) and the internal energy (ΔE_{in}) so that there is almost no differences in the binding energy between Fpg/FG and Fpg/G before solvation effects are considered. Hence Fpg stabilises FapydG by 8.4 $kcal/mol$ more than G, mainly due to the solvation free energy (ΔG_{pol}), which is reasonable since water molecules are required to establish indirect interactions between Fpg and the lesion as shown in figure 3.2. The FapydG residue that is favourable in binding to Fpg may also explain the rigidity when the lesion bound to the binding pocket in figure 3.6. Finally, Fpg is likely to be able to discriminate the lesion from the non-lesion once the nucleobase is extrahelical.

Table 3.2: Relative binding free energy of Fpg/FG and Fpg/G complexes from MM/GBSA approach

Complex	ΔE_{vdW} (kcal/mol)	ΔE_{es} (kcal/mol)	ΔE_{in} (kcal/mol)	ΔG_{pol} (kcal/mol)	ΔG_{nonpol} (kcal/mol)	$\Delta G_{binding}$ (kcal/mol)
Fpg/FG	-27.0±2.0	-8.8±4.8	2.7±1.0	9.4±3.8	-3.2±0.1	-26.9±2.7
Fpg/G	-22.9±2.0	-17.4±7.6	6.7±1.6	17.8±4.7	-2.8±0.1	-18.5±3.8

3.4.2 Specific interactions of FapydG versus G

Initial structures of Fpg/FG and Fpg/G complexes after minimisation and equilibration, focusing on interactions and distances between some active residues and the ligand (FapydG or G), are shown in figure 3.7. FapydG differs from guanine in hydrogen bonding capability by having an acceptor O, a donor N6 and conversion from an acceptor N7 (guanine) to a donor N5 (FapydG). Since a water molecule which links between the carbonyl O atom of the formamide group and Tyr238 disappears after equilibration and Tyr238 has a very close contact to the phosphate backbone, it is arguable that the water-assisted interaction between the formamide group and Tyr238, as suggested by Coste *et al.* [54], is entirely labile (in figure 3.2).

As demonstrated in figure 3.7(a), the distinct formyl group of FapydG is presumably recognised by potential interactions with Arg220 through hydrogen bonding with either the amide backbone (3.06 Å) or the secondary amine side chain (3.23 Å). Hydrogen bond donors at N5 and N6 positions of FapydG are able to efficiently establish a hydrogen bond network via either direct or indirect interactions with Met75 and Glu76, where the N7 atom of guanine is indirectly bonded to the carbonyl group of Glu76 via a water molecule shown in figure 3.7(b). Additionally, a common functional group of FapydG and guanine, the carbonyl O4 (FapydG) or O6 (guanine) atom, is apparently hydrogen bonded by the different main chain of Ile219 and Arg220, respectively.

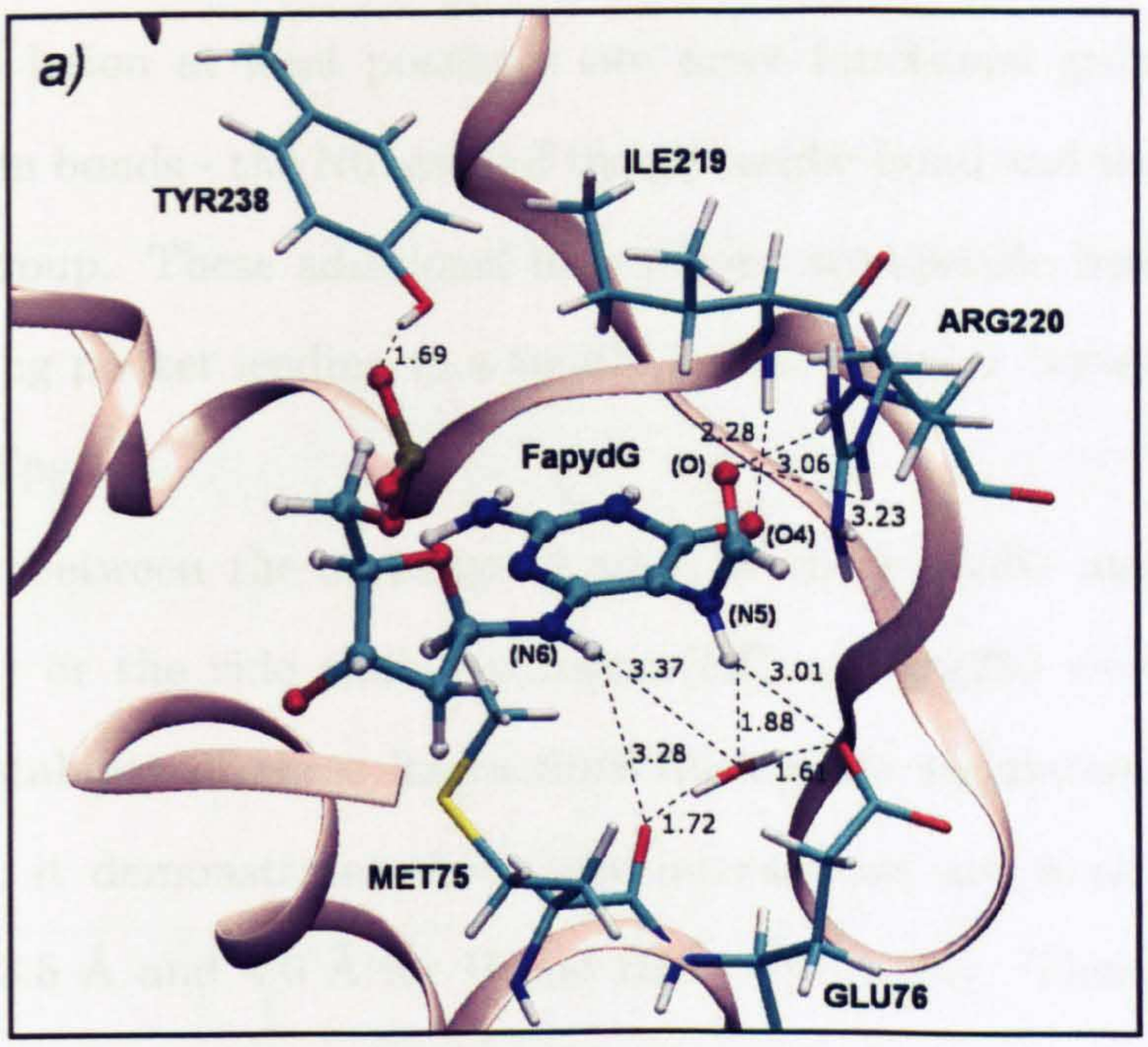


Figure 3.7: Comparisons of conformation of the recognition binding region containing FapydG (in panel a)or guanine (in panel b), Met75, Glu76, Ile219, Arg220 and Tyr238. The nucleotides are shown in small ball and stick, the protein backbone in a ribbon diagram and the surrounding residues in a licorice model. The favourable interactions between the nucleotide, water and amino acids are indicated by *black dashed lines* associated with the distance.

According to the comparison of the structures above, it reveals that the hydrogen bonding patterns of Fpg/FG and Fpg/G complexes are obviously different. The FapydG lesion at least possesses two more functional groups available to form hydrogen bonds - the N6 atom of the glycosidic bond and the O atom of the formamide group. These additional interactions are specific between the lesion and its binding pocket leading to a tightly bound complex between the FapydG residue and Fpg.

Distances between the carbonyl O atom of the FapydG and the backbone hydrogen (H) or the side chain hydrogen (HE) of Arg220 were calculated to explore the stability of these interactions during the simulation. As shown in figure 3.8(a), it demonstrates that these interactions are weak with averaged distances of 3.5 Å and 4.0 Å for H and HE, respectively. Thus, an alternative contact such as the water-mediated interaction may be required to maintain the protein-lesion complex. Figure 3.8(b) demonstrates a snapshot representing a hydrogen bonding network between the carbonyl O atom and Arg220 through a water molecule. On the other hand, the guanine has no functional groups to interact with Arg220.

3.5 Conclusions

To understand mechanisms of DNA damage recognition at an atomic level, x-ray crystallographic structures remain an important source of information to begin with. The most informative structure of the Fpg protein bound to a FapydG-containing duplex was selected as an initial model. Subsequently a model of the Fpg-DNA duplex containing an extrahelical guanine was also constructed to represent a version of the complex when Fpg flips the normal nucleobase out from the double helix. Prior to performing MD simulations, parameters for Fpg residues were carefully amended in terms of their protonation states,the pre-catalytic stage of the N-terminal proline, and the geometry of the zinc finger motif attempting to produce a most reliable model.

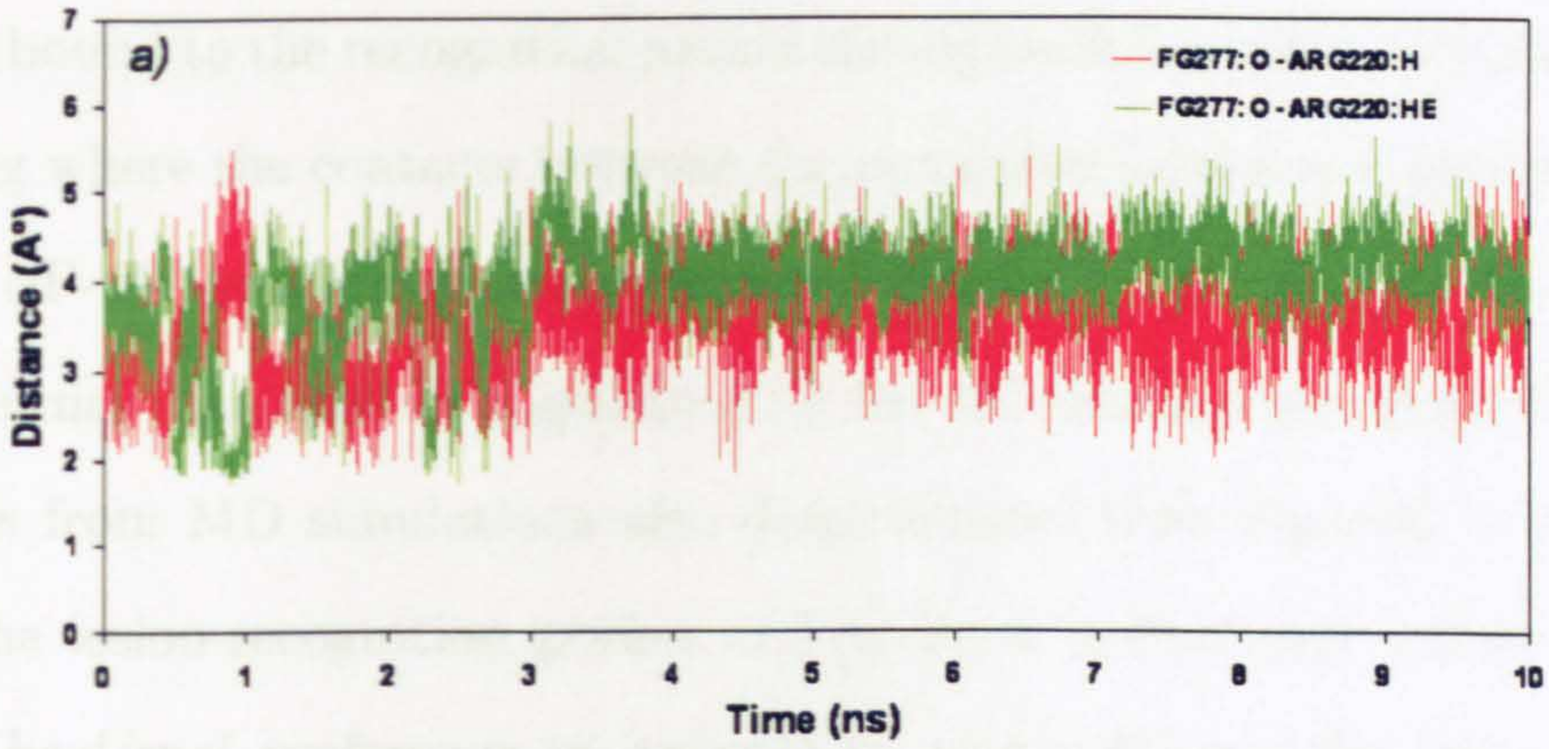


Figure 3.8: (a) Distances in Å corresponding to hydrogen bonds between the carbonyl O atom of the FapydG (FG277:O) and the backbone hydrogen (ARG220:H) or the side chain hydrogen (ARG220:HE) of Arg220 during the simulations. (b) A snapshot of the Fpg/FG complex indicates possible direct and indirect interactions between the carbonyl group of the FapydG residue and either the main chain or the side chain of Arg220.

Molecular modelling and MD simulation studies were clearly evident that if Fpg distinguishes the lesion by the extrusion of each base into the recognition pocket, the repair protein is capable of discrimination of the lesion from the non-lesion once the base is extrahelical. The present results have shown that FapydG is tightly bound to the recognition pocket through either direct or indirect hydrogen bonding where the contacts between the nonplanar formamide group with a part of the αF-β9 loop are an important component. The αF-β9 loop of the Fpg enzyme may function as a gatekeeping for the damage recognition. Energetic analysis from MD simulations also demonstrated that FapydG is preferable to G in the lesion-recognition pocket of Fpg by $\approx$ 8 *kcal/mol* which corresponds to a 7 *kcal/mol* preference to insert 8OG versus G into the lesion-recognition pocket of hOGG1 [134]. It was finally suggested that Fpg is able to discriminate the subtle structural changes of the FapydG lesion from its normal counterpart by producing a stable complex of the lesion and Fpg while denying the normal nucleobase.

Chapter 4

Base Flipping

4.1 Introduction

4.1.1 Base flipping and biological relevance

Since the first base flipping event was observed in a co-crystal ternary complex of cytosine-5-methyltransferase, its DNA substrate, and *S*-adenosylhomocysteine in 1994 [135], base flipping was then termed as a natural phenomenon of DNA where a target base is completely rotated out from a stacked position inside the double helix into an extrahelical location. Such a process appears to be a general and pivotal mechanism for DNA modification [135, 136] and for DNA repair proteins to deal with the target bases [137, 138]. However, DNA-modifying enzymes such as DNA methyltransferases are sequence-specific binding proteins that flip a normal base out of the helix [139], whilst DNA repair enzymes are sequence-independent that must recognise and flip damaged bases out of the helix no matter what the sequence is. Nonetheless, it has been suggested that mechanisms of DNA repair proteins in order to detect and repair their targets are likely different from those of sequence-specific enzymes [140]. Base flipping may also occur in the early stages of DNA strand separation and unwinding as would be required during the DNA replication and transcription [141, 142].

4.1.2 Mechanisms of base flipping

There are two possibly mechanisms of base flipping [141]. The first is a protein-facilitated mechanism in which the protein recognises the intrahelical target. A proposed flipping mechanism called a "pinch-push-pull" model is used by Uracil-DNA glycosylase (UDG), the prototypical glycosylase that removes uracil from double-stranded DNA [140]. The pinch-push-pull mechanism is initiated by the protein searching for the damage by slightly bending the DNA (pinch), then inserting some appropriate residues to push the base out via the major groove (push) and finally capturing the base inside its specific binding pocket by attraction of the complementary active site residues (pull). It has been believed that the pinch-push-pull method is a common mechanism used by other related glycosylase enzymes [140].

Alternatively, a rarely flipping event is spontaneous opening of a nucleobase from helical DNA, or DNA breathing, where the flipped-out base is recognised by the protein. The disruption of hydrogen bonds between the complementary bases results in the occurrence of base-pair breathing and flipping of the base out of the helical stack. Spontaneous base pair-opening can be indirectly examined from exchanging rates of labile base protons with the surrounding solvent [143]. Using NMR imino proton exchange measurements, it was demonstrated that an opening process of a natural Watson-Crick (WC) base pair occurs in the millisecond range and it is a sequence-dependent property [144, 145, 146]. It also revealed that base flipping of A:T and G:C base pairs takes 1-10 and 5-50 ms, respectively, with remarkably high energy of activation $\approx$ 20-80 $kcal/mol$ while bases remain in the flipped-out state for 10-100 ns [147]. In DNA repair, the uracil opening, for instance, was observed in the MD simulations of DNA containing a G:U wobble base-pair [148]. It was also evident that UDG searched the target by capturing an extrahelical uracil [149].

Although extensive studies have been carried out for base flipping mechanisms by DNA glycosylases, even in the well-studied UDG, questions of whether repair

proteins presumably locate a particular damaged base via its chemical specificity and/or the intrinsic curvature of damaged DNA; or may use the benefits of spontaneous damaged-base flipping to be an initial lesion recognition remain unclear. It was anticipated that differences in free energy pathways associated with the conformational changes for damaged base opening relative to its normal counterpart can explain how base flipping plays an important role in DNA damage recognition. Unfortunately, experimental studies are unable to define the energetics and the structural transformation of base flipping, whereas the long time scale of the flipping process and the high intrinsic energy barrier make it inaccessible by conventional MD simulations.

4.1.3 Umbrella sampling and WHAM analysis

When two configurations of interest of a molecular system are separated by a large energy barrier, standard MD simulations which are initiated from one configuration are unlikely to overcome the high potential barrier leading to poor sampling or even unsampling the configuration space of another configuration. One approach to overcome the problem is an umbrella sampling technique. The umbrella sampling approach which was first introduced by Torrie and Valleau in 1977 [150] attempts to improve the sampling of such system by addition of biasing potentials to maintain intermediate configurations in high energy regions between two configurations. In other words, high energy configurations can be sampled.

An umbrella sampling method is implemented within the *sander* module of the AMBER suite enabling one to simulate the microscopic system in the presence of a harmonic biasing potential. The biasing potential typically applies to distance, angle, or dihedral variables (known as reaction coordinates or RCs). The biased simulation of an initial coordinate of interest (often called a "window") offers the distribution of values of this coordinate during the simulation. The biased simulations are then repeatedly performed with different ranges of

the coordinate moving towards the target configuration. It is crucial that each simulation or window must be equilibrated and the probability distribution of each biased simulation must overlap with that of its neighbouring windows.

All equilibrium simulations are then transferred to the Weight Histogram Analysis Method (WHAM) program written by Alan Grossfield [151] based on the WHAM approach by Kumar *et al.* [152]. The WHAM approach is employed to calculate a potential of mean force (PMF), which describes the change in free energy (ΔG) along the chosen RC, from unbiased probability distribution ($W(x)$) and temperature (T) in each equilibrium window via equation 4.1.

$$\Delta G = -k_B T \ln W(x) \tag{4.1}$$

where the k_B is the Boltzmann constant and the $W(x)$ is corrected by removing the effect of the biasing potentials from the simulations. The WHAM procedure is suggested to yield the best estimates of free energies since it takes into account all the simulations so as to minimise the statistical errors and can optimise the overlaps of $W(x)$ from different windows [153]. PMFs are subsequently obtained by connecting the free energy along the coordinate. Construction of PMFs using the umbrella sampling combined with WHAM method has been frequently used to understand biomolecule conformational changes at an atomic level such as base flipping in DNA [154, 155] or in RNA [156], and conformational changes in an allosteric binding site [157].

Umbrella sampling MD simulations of base flipping pathways in *B*-DNA have been performed by applying restraints on a variety of RCs. Studies of the opening of the WC A:T and G:C base pairs using an angle restraint to generate the energy profiles have revealed that the pyrimidine bases are flipped towards major and minor groove pathways with similar energy profiles while the purine bases are rotated through the more favourable major groove [158, 154]. However, the deformability of DNA backbone was ignored from the angle restraint approach. One novel method that took the backbone flexibility into account was the use of

umbrella sampling with a centre-of-mass (COM) dihedral RC to calculate PMF for G:C base pair opening in the absence of protein [155] and in the presence of cytosine-5-methyltransferase [159]. A schematic diagram of the COM dihedral for the cytosine flipping is presented in figure 4.1. The COM dihedral is calculated from four sets of heavy atoms: (i) the guanine and cytosine bases forming a base-pair at the 3' adjacent to the target cytosine, (ii) the sugar moiety attached to the 3' guanine, (iii) the sugar moiety attached to the flipping cytosine base, and (iv) the flipping cytosine base [155]. This method offered comparable results for guanine and cytosine flipping pathways to the previous angle restraint reports when the protein is absent. However, in the presence of protein the flipping of the target cytosine base occurs through the preferable major groove of the DNA and the process is facilitated by the protein [159].

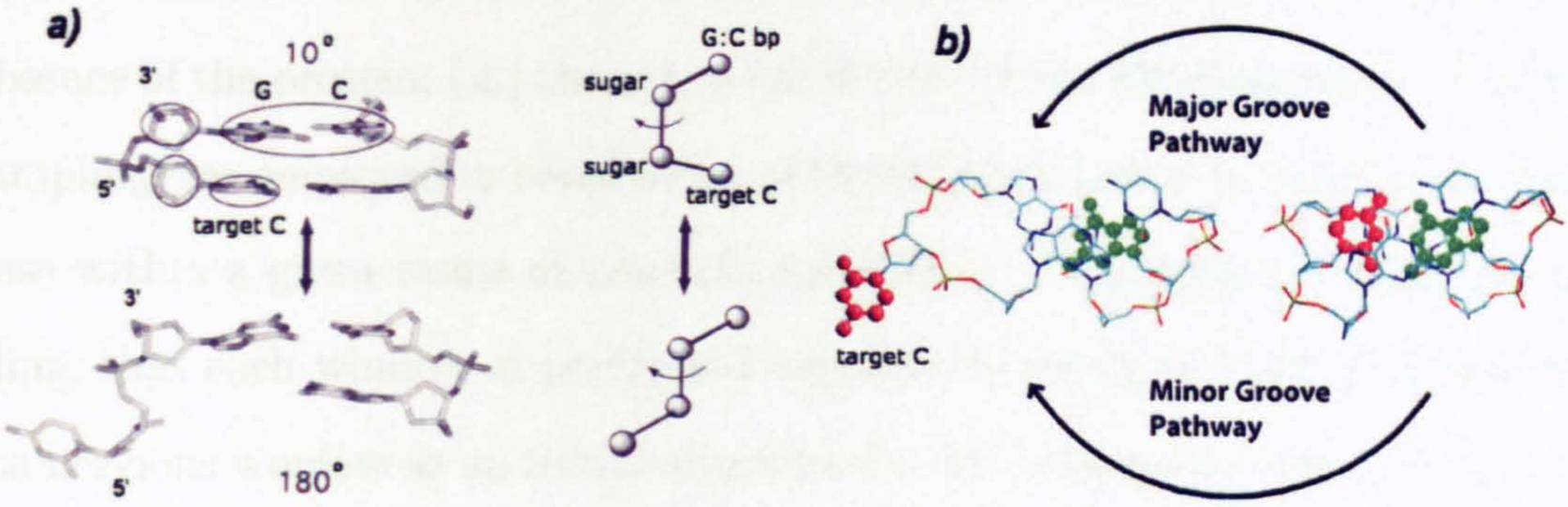


Figure 4.1: (a) A schematic diagram of the centre-of-mass dihedral constraint for cytosine flipping (see text), the intrahelical C conformation for COM dihedral value ≈10° (top) and the flipped-out C in ≈180° (bottom). A simplified diagram of COM dihedral is shown on the right. (b) Flipping of the target C base (red) from the helical conformation (right) to a flipped conformation (left) via the major or minor grooves. The figures are adapted from [155, 159].

4.1.4 Aims and objectives

Base flipping plays a key role in the base excision repair of DNA base damage since all lesions must be rotated out from the helical duplex in order to be cleaved from DNA by DNA glycosylases. To understand how the proteins recognise the lesion over its normal equivalent through the flipping mechanism, it is necessary to

investigate their structural transformation and energetic changes involved in the process. Moreover, crystallographic structures of the repair protein-DNA systems have revealed slightly or sharply bent DNA conformations bound to the active binding pocket of repair proteins, as shown in table 2.1. It was also suggested by theoretical molecular modelling that base pair opening is coupled with DNA bending [160]. Thus, the influence of DNA deformation on the base flipping was also taking into account for this study.

The aim of the work in this chapter was to generate profiles of the relative free energy changes of the FapydG or guanine base flipping pathway in three distinct circumstances of the DNA: (i) the distorted DNA structure in the presence of the Fpg protein as being observed in the crystallographic structure (PDB entry 1XC8) [54]; (ii) the sharply kinked conformation of the DNA duplex d(TCTTTXTTTCTC)·d(GAGAAACAAAGA) where X is FapydG or G in the absence of the protein; (iii) the canonical B form of the DNA sequence. Umbrella sampling was employed in conjunction with MD simulations to restrain the target base within a given range of reaction coordinates. Conventional umbrella sampling, that each window is performed sequentially using the last structure from the previous window as an initial structure for the subsequent simulation, is very time-consuming. A targeted MD approach was therefore used to force the flipped base to swing back into the helical duplex. A set of coordinates along the flipping pathway from the flipped state to the intrahelical position were consequently generated and used as an initial structure for each window. Finally, by a means of the combined method, the simulations were performed in parallel processing and structural intermediates associated with energetic changes estimated using the WHAM approach.

4.2 Methods

4.2.1 Construction of pre-flipped models

Two actual states of FapydG, intrahelical and extrahelical conformations, are essential in order to investigate the flipping pathway. Unfortunately, only the crystal complex of the *Ll*Fpg with DNA containing the extrahelical *c*FapydG base (PDB entry 1 XC8) [54] is available. Molecular modelling techniques therefore were used to construct a pre-flipped structure of FapydG-containing DNA bound to Fpg. The intrahelical FapydG was built with respect to the previous studies in which the FapydG residue exhibits the *anti*-glycosidic conformation with a nonplanar formamide group (see [93] and section 2.1.2). The pre-flipped DNA model was basically built by rotating α, β, γ and δ torsion angles of the backbone of the damaged nucleotide into the duplex and the χ torsion of FapydG and the opposite cytosine to re-establish the hydrogen bonds (figure 4.2), while the remaining residues and the kinked conformation were retained as within the original 1XC8 model. The extrahelical guanine DNA model was constructed using the *Leap* module with a similar way as described in section 3.2.2, and the intrahelical model was built as described above.

Fitting the intrahelical-FapydG DNA model into the binding pocket of the *Ll*Fpg protein, steric clashes between the FapydG residue and the intercalating triad from the N-terminus (Met75, Arg109, and Phe111) were found. The Fpg model was therefore initially remodelled with respect to the *apo*-form of *Tt*Fpg (PDB entry 1EE8) [121] due to the absence of structures of the uncomplexed enzyme from the same source. Between the free *Tt*Fpg and the bound *Ll*Fpg, there are slightly different orientations mainly in the hinge region and the N-terminal domain, as illustrated in figure 4.3. Electrostatic potential surfaces of the *apo*-form of *Tt*Fpg compared to the DNA-bound form of *Ll*Fpg with the removal of the DNA were generated in VMD by an Adaptive Poisson-Boltzmann Solver (APBS) approach [57]. The figure demonstrated the highly positive DNA-

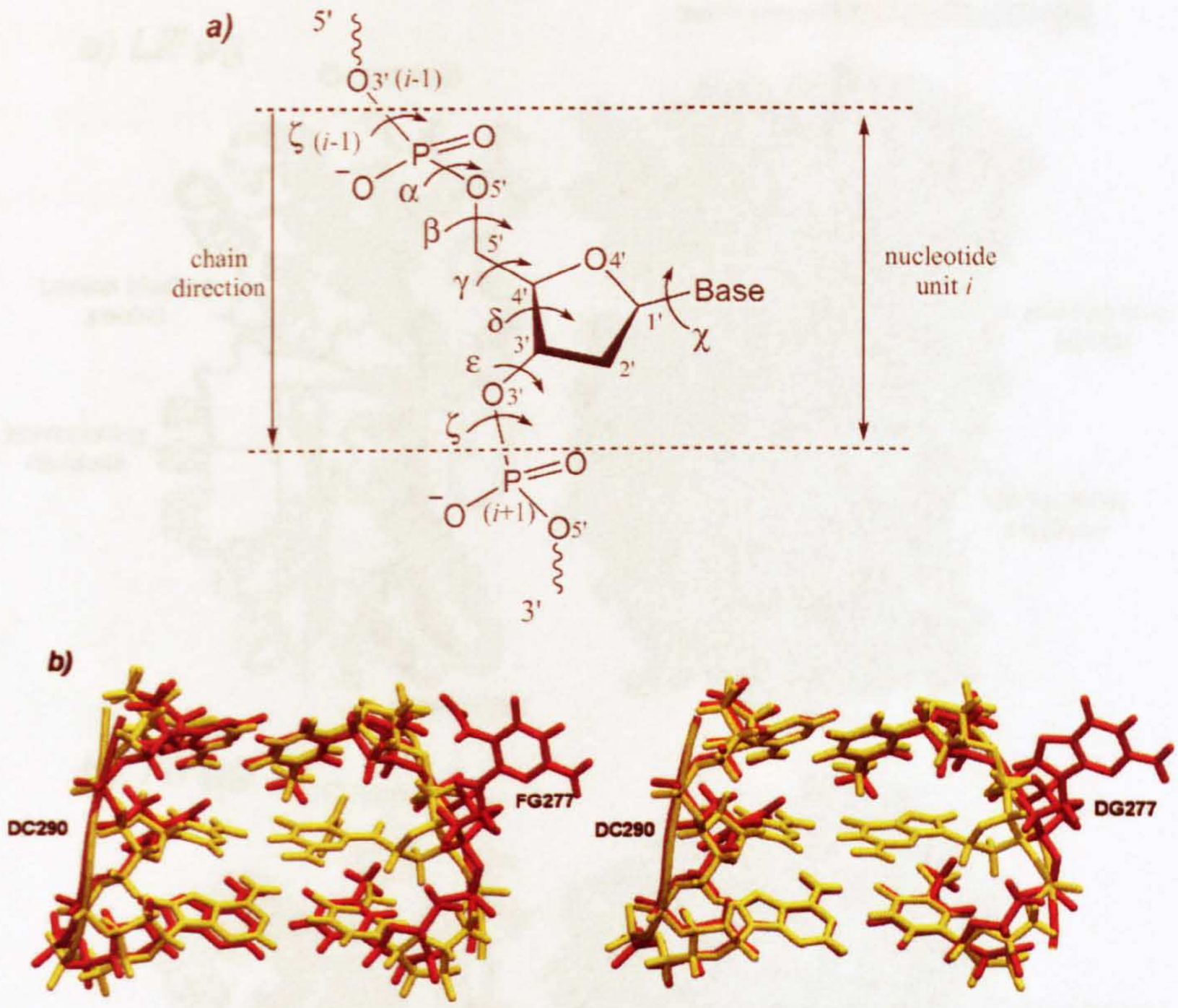


Figure 4.2: (a) The notation for torsion angles in a polynucleotide backbone. Each curved arrow demonstrates a rotatable bond. (b) Starting structures of the intrahelical DNA model in yellow and the extrahelical model in red showing that only the FG:C and G:C base pairs are relocated while the other parts are retained including the distorted conformation of DNA. The picture illustrates the DNA tribase of the damaged (left) and the undamaged (right) DNA models.

binding cleft of the enzyme with a relatively tightly lesion-recognition pocket of the "closed" state of *Ll*Fpg when binds to the damaged DNA. Cavity calculations of the lesion binding pocket by the VOIDOO program [161] indicated a larger space of the *Tt*Fpg pocket (70.2 Å^3) compared to that of the *Ll*Fpg (41.4 Å^3). It was also noticeable that the intercalating residues in the *apo*-form of *Tt*Fpg are slightly more open relative to those of the closed *Ll*Fpg conformation. More details of the structure comparison of the *apo*-form of *Tt*Fpg and the *Ll*Fpg-DNA complex can be seen in [56]. Homology modelling by the web-based SWISS-MODEL workspace [162] was subsequently used to generate an *apo*-form of *Ll*Fpg. Unfortunately, it failed to resolve the steric clashes between FapydG and those of the intercalating triad.

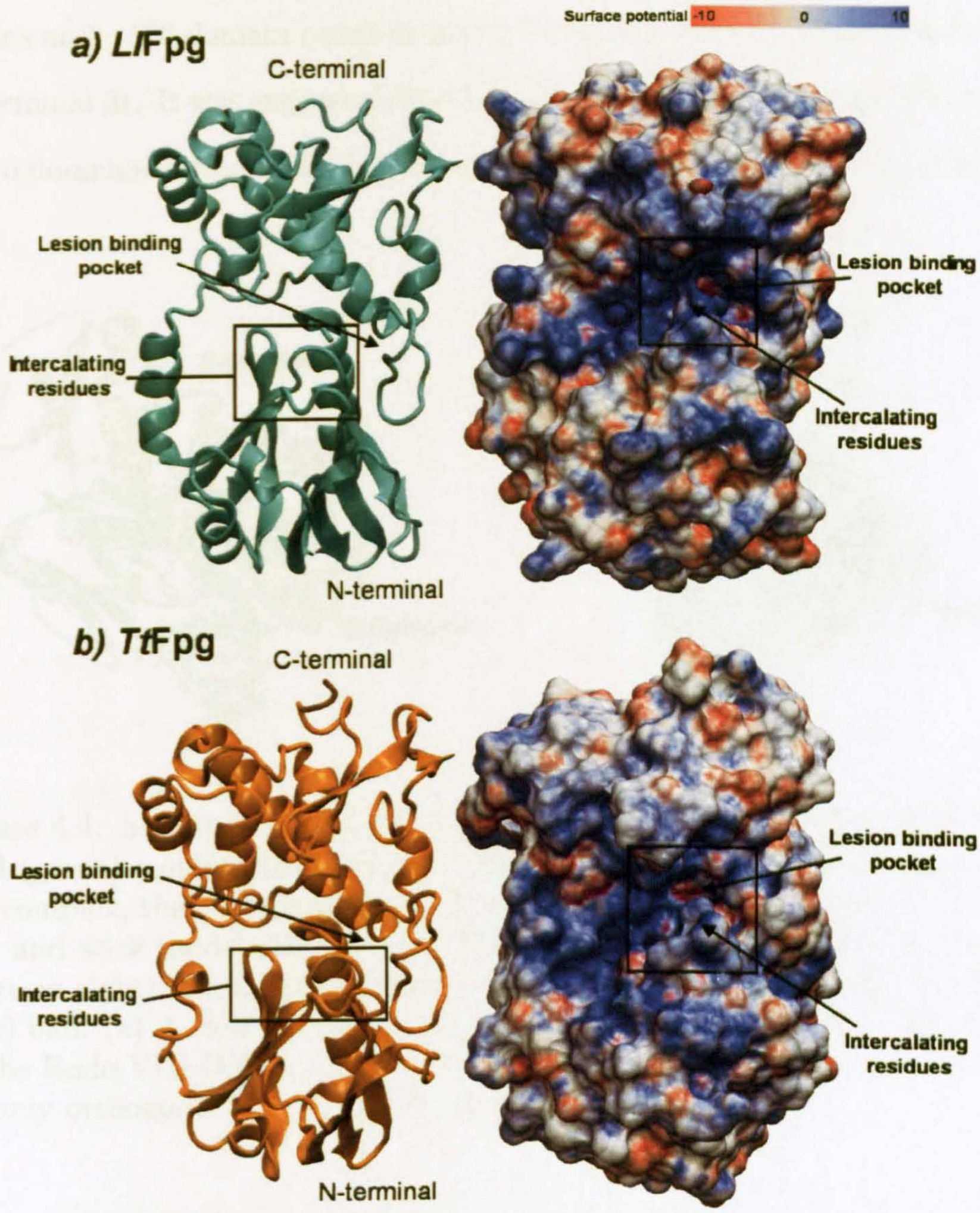


Figure 4.3: Structures of (a) the bound *Ll*Fpg and (b) the free *Tt*Fpg are represented by a cartoon diagram and a solvent-accessible surface model. The solvent-accessible surface models are coloured according to electrostatic potential (positive in blue, negative in red and neutral in gray) demonstrating the lesion-recognition binding pocket and the intercalating residues.

Another close related source for the bacterial Fpg enzyme is the bacterial endonuclease VIII (Endo VIII or Nei) which belongs to the same H2TH structural family [58] and possesses both *N*-glycosylase and AP lyase activities as with Fpg. Since the free and the DNA bound Endo VIII are structurally characterised, these demonstrate two distinct states of the enzyme in the presence and absence of DNA as shown in figure 4.4 [163]. A close structural similarity between the C-terminal

domains and a 50° domain rotation of the N-terminal domain were observed after a C-terminal fit. It was suggested that DNA-binding may induce the movement of the two domains of Endo VIII into the closed state due to the flexible interdomain hinge.

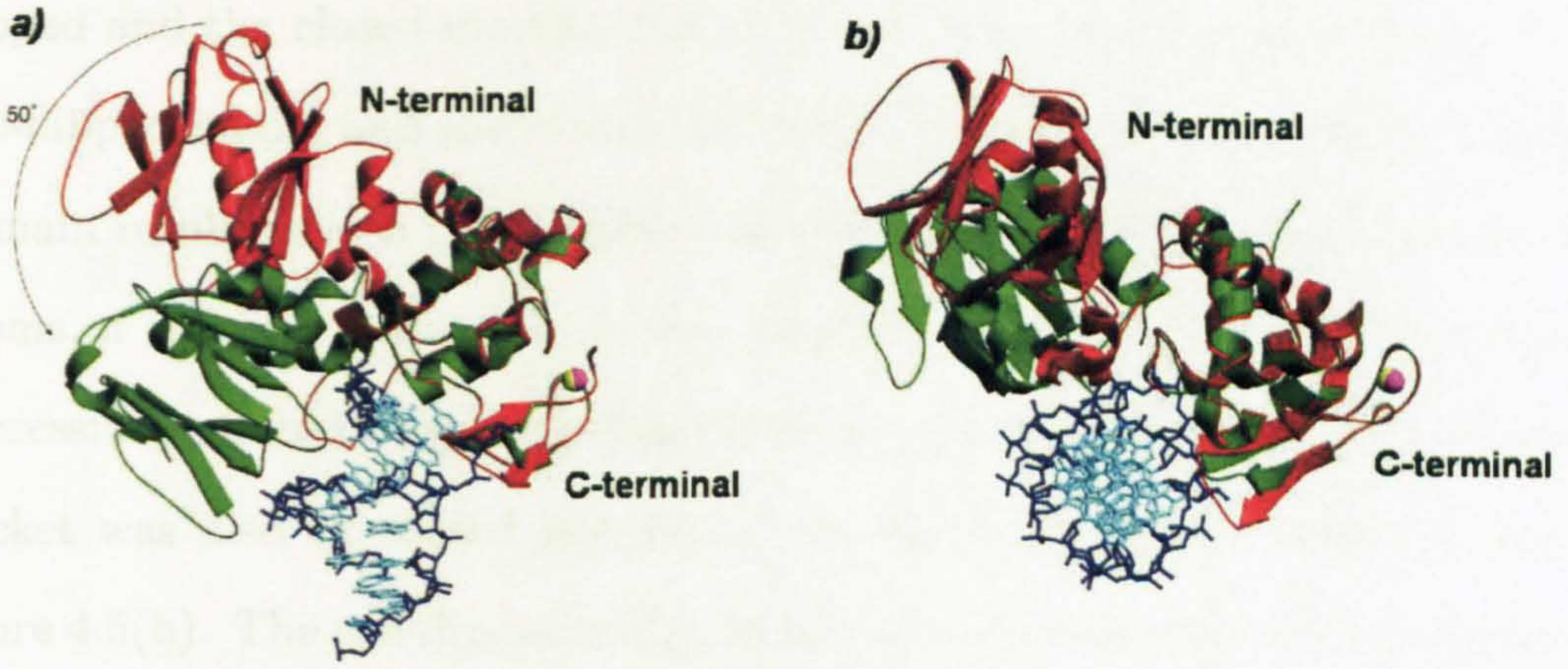


Figure 4.4: Superposition of the free Endo VIII (red) and the bound Endo VIII (green) resulted from a least-squares fit of the C-terminal domains. In the complex, the enzyme is shown in a cartoon diagram, the DNA in a small ball and stick model (backbone in blue; bases in cyan), and the zinc atom (bottom right) in a magenta (the free enzyme) or yellow (the complex structure) ball. (a) A view perpendicular to the long axis of the Endo VIII protein in the Endo VIII-DNA complex. (b) The same superposition model approximately orthogonal to (a). The figure is taken from [163].

Based on the finding of the DNA-induced conformational changes in Endo VIII, the damage recognition by Fpg may be initiated by a similar approach in which the enzyme binds to damaged DNA followed by the closing movement of the domains upon DNA binding. Thus, to construct an initial pre-flipped model of *Ll*Fpg without the steric clashes, the N-terminal domain of *Ll*Fpg which contains the intercalating residues (Met75, Arg109 and Phe111) was shifted away from the DNA minor groove with respect to that of the *apo-Tt*Fpg protein while the C-terminus of the closed *Ll*Fpg conformation was retained as in the crystal structure.

Superposition of the pre-flipped model and the bound *Ll*Fpg-DNA complex resulted in a relatively high RMSD value of 2.42 Å (backbone atoms of residues

5 to 265). On the other hand, the individual domain overlaps resulted in an RMSD value of 0.78 Å for the C-terminal domain (backbone atoms of residues 150 to 265), and that of the N-terminal domains showed an RMSD value of 0.66 Å (backbone atoms of residues 5 to 115). These RMSDs obviously demonstrated that the structures of the individual domains were nearly equivalent in the pre-flipped and the closed models. Figure 4.5(a) illustrates the superposition of the pre-flipped model and the bound *Ll*Fpg-DNA complex based on the C-terminal domain resulting in a large RMSD value of 4.69 Å of the N-terminus (backbone atoms of residues 5 to 115). It was suggested that the N-terminal domain was successfully shifted away from that of the crystal structure and the lesion binding pocket was also expanded particularly in the intercalating region as shown in figure 4.5(b). The pre-flipped *Ll*Fpg model consequently permits the intrahelical-FapydG DNA model to nicely fit inside the binding pocket of the protein. A pre-flipped model of the *Ll*Fpg/G complex was also constructed similar to the method described in section 3.2.2. Both pre-flipped models of the Fpg/FG and Fpg/G complexes were energy-minimised to get rid of any bad contacts in the protein-DNA complexes.

4.2.2 Production of flipping trajectories

After completing the construction of the pre-flipped models of both Fpg/FG and Fpg/G complexes, the base flipping pathway was able to be approximated using targeted molecular dynamics (TMD). TMD is a method to simulate a conformational transition pathway between two known structures at a given temperature by applying a time-dependent, geometrical constraint [164]. The TMD method which is available within the *sander* module of the AMBER suite generally adds positional restraints to attain a range of target structures within a given mass-weighted RMSD tolerance of a reference structure from an initial structure.

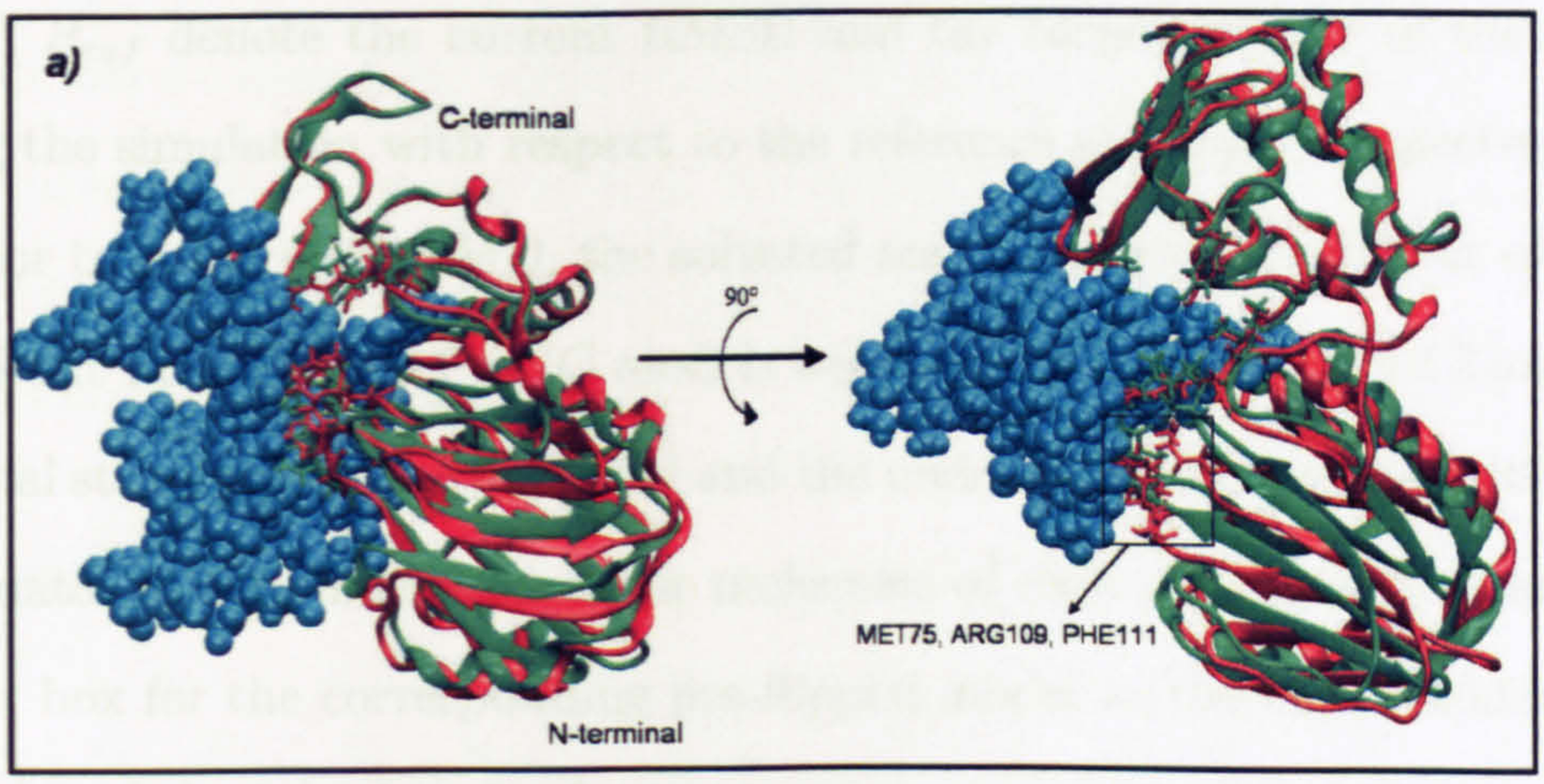


Figure 4.5: (a) Superposition of the Fpg proteins with respect to the C-terminal domain is represented by a cartoon diagram in which the closed *Ll*Fpg conformation in green and the pre-flipped state in red. The extrahelical FapydG-containing DNA is shown by a cyan van der Waals model. Some active residues inside the binding pocket are shown by a licorice model. (b) Shifting away of the intercalating triad (Met75, Arg109 and Phe111) showing in red from the closed *Ll*Fpg structure in green.

The additional restrained energy inserted on the selected target structures (E) is calculated by

$$E = \frac{kN}{2}(R_i - R_{ref})^2 \tag{4.2}$$

where k is a user-defined force constant, N is the number of the target atoms, and R_i and R_{ref} denote the current RMSD and the target RMSD of the structure during the simulation with respect to the reference structure, respectively.

Prior to performing TMD, the solvated and minimised structures of the fully flipped-out Fpg/FG and Fpg/G models were taken from section 3.2.3 and used as an initial structure for the damaged and the undamaged models respectively. The coordinates of the minimised water molecules of each system were then used as a water box for the corresponding pre-flipped model as the same atomic number and order of the initial and reference systems are required in order to carry out TMD. Fortunately, there was no steric clashes between the Fpg-DNA complex and the water molecules. The base flipping process was simulated using the TMD approach to force the extrahelical base back into the helix as well as the opening of the N-terminus by defining all C_α atoms of the Fpg protein and the FapydG:C or G:C base pairs as the target structure. A force constant of 5.0 $kcal/mol/Å^2$ was used in TMD and the weight change technique was combined to continuously reduce the target RMSD from 4.20 Å to zero with respect to the reference structure within 1-ns simulations. In all simulations, the terminal base pairs of the duplex were harmonically restrained using positional restraints with a force constant of 3.0 $kcal/mol/Å^2$. The flipping trajectories eventually were generated to provide starting structures for each reaction coordinate (RC).

Since the COM dihedral constraint was first introduced in the base flipping study using the program CHARMM (Chemistry at Harvard Macromolecular Mechanics) in 2002, it has been suggested to be a novel RC for base flipping studies [155]. Unfortunately, extensive modification for the source code of the *sander* program is required to apply this approach in the calculation of MD simulations. In fact, a relatively high biased potential is also required to restrain those

groups of atoms, the RC chosen in this study was the θ dihedral angle spanned by the atoms FG277:C6 or DG277:C4 , FG277:C4' or DG277:C4' , DT278:C4', and DT278:N3, where FG denotes the FapydG nucleotide (figure 4.6). The flipping dihedral angle θ was calculated compared to the COM dihedral angle over the targeted MD trajectory of the Fpg/FG complex using the *ptraj* program. Figure 4.7 exhibits a comparable value between the θ angle and the COM dihedral angle. The trajectories also revealed that the minor groove flipping as increasing from 0° to 180° is preferred in this recognition process.

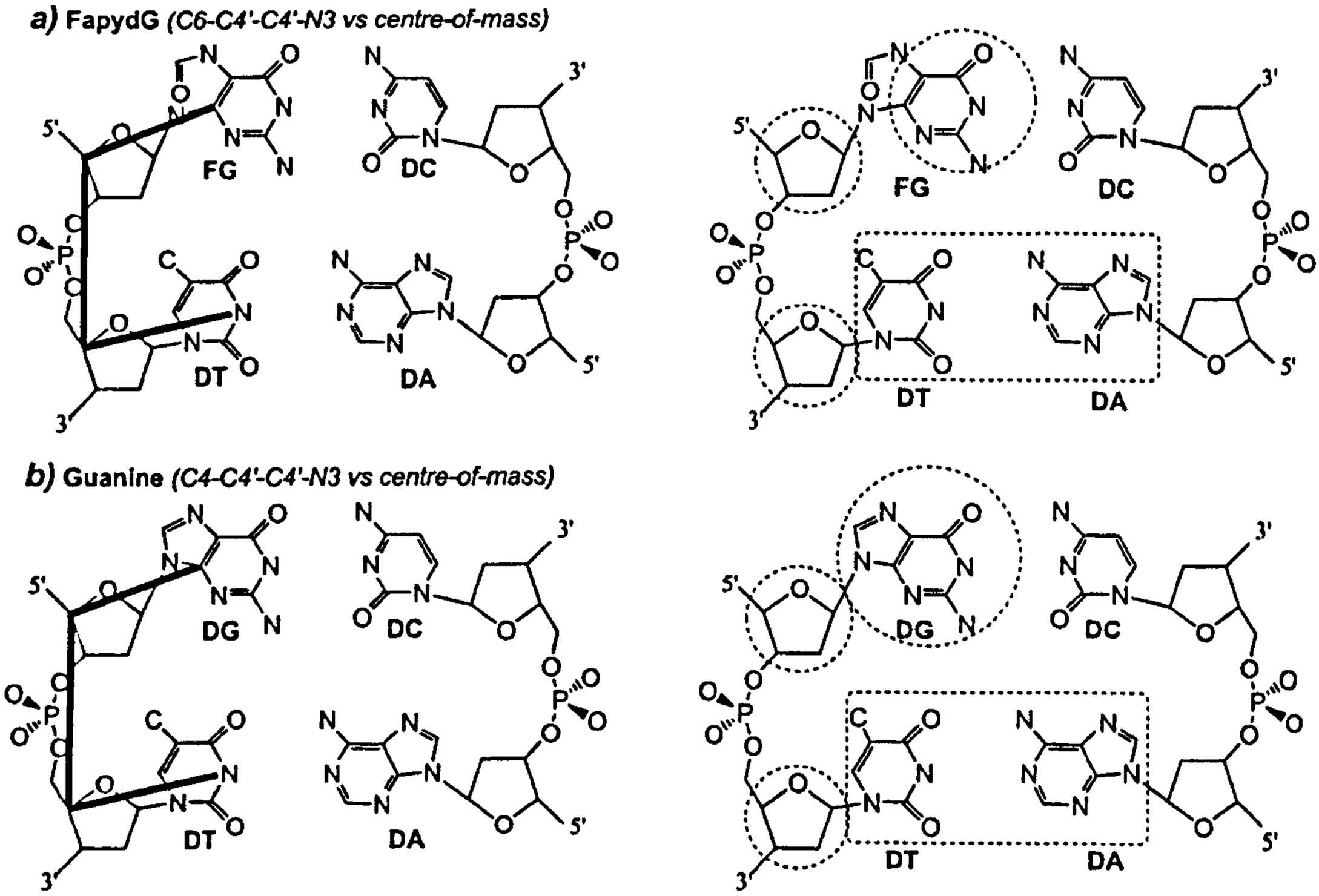


Figure 4.6: Depiction of the dihedral angle reaction coordinate θ to study base flipping are illustrated by a *bold line* between C6 and C4' atoms of FG or C4 and C4' atoms of DG and C4' and N3 atoms of 3' DT, compared to the COM dihedral angle defined by four sets of atoms used to calculate the flipping angle in the *ptraj* program.

Targeted MD was also used to generate the trajectories of the canonical *B*-DNA containing FapydG or G when the protein was omitted. All parameters for the Fpg protein were retained as described in section 3.2.1. The Fpg-DNA systems were neutralised, solvated and equilibrated as described in section 3.2.3 while the protein-free simulations were performed as described in section 1.4.2.

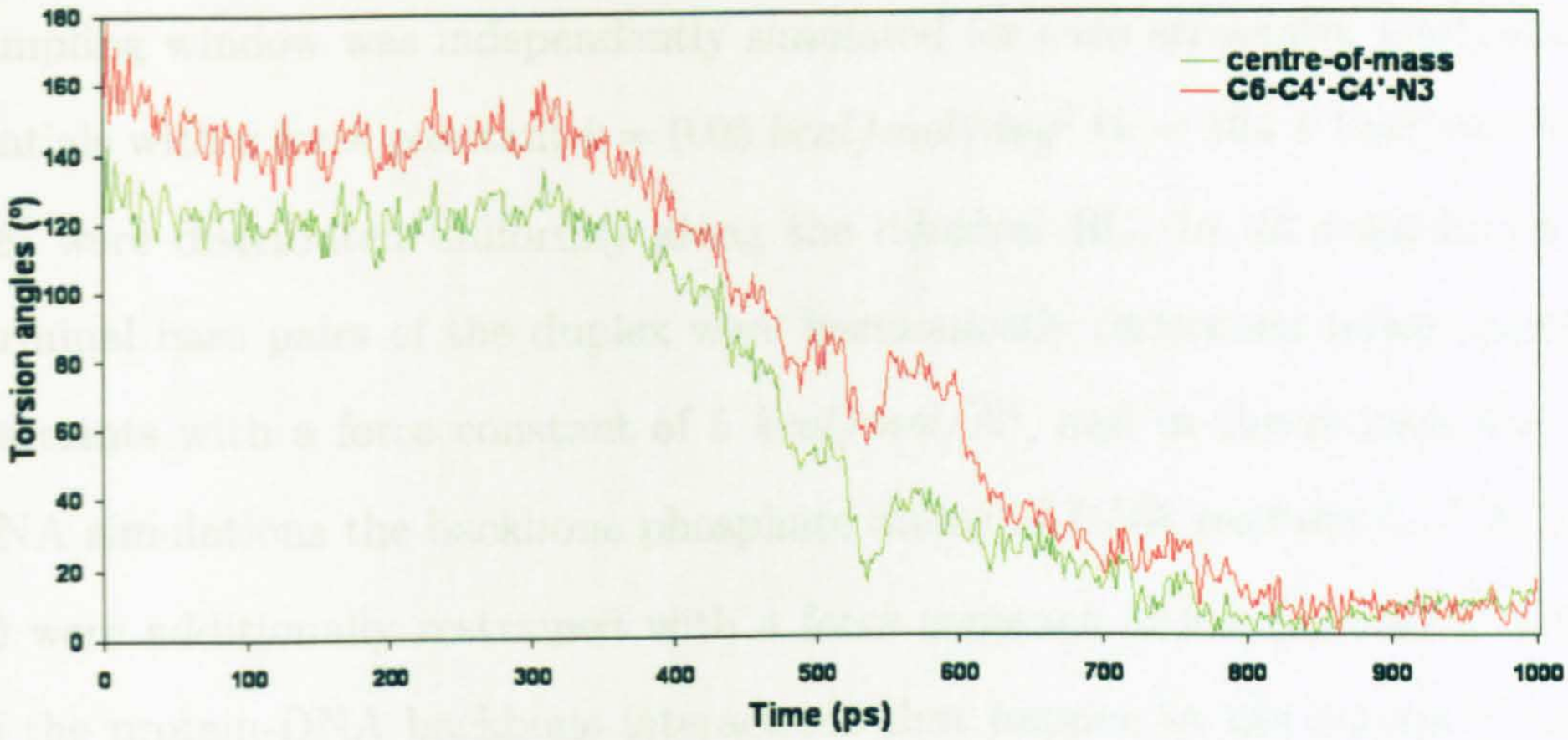


Figure 4.7: Dihedral angle plots of the flipping FapydG base over the targeted MD defined in 2 different approach, the θ and the COM dihedral angles.

4.2.3 Umbrella sampling and PMF Calculations

The umbrella sampling method was used to calculate the potential of mean force (PMF) as a function of the θ dihedral angle for the damaged (C6-C4'-C4'-N3 for FapydG) and the undamaged (C4-C4'-C4'-N3 for G) models (see figure 4.6) in three different environments - the Fpg-DNA complex, the protein-free distorted DNA duplex d(TCTTTXTTTCTC)·d(GAGAAACAAAGA) where X is FapydG or G, and the canonical B form. The starting structures for each umbrella simulation of the Fpg-DNA and the *B*-DNA systems were taken from the targeted MD trajectories when the initial structures for the bent duplex were obtained by removal of the protein structure from the structures of the Fpg-DNA complex. Structures with the θ angle from 0° to 180° in 5° increments were selected resulting in 37 flipped orientations of the base through the minor groove for each studied system. The major groove flipping pathway as decreasing from 0° to -180° of the θ angle was omitted in this study as the visual inspection of the Fpg-DNA crystal complex revealed that the major groove of the bent duplex is compressed and packed with the C-terminal domain of the Fpg protein whilst widening of the minor groove is observed.

These initial structures were equilibrated for 0.2 ns, and one 1-ns umbrella sampling window was independently simulated for each structure. Umbrella potentials with a force constant $k = 0.05\ kcal/mol/deg^2$ ($k = 164.1\ kcal/mol/rad^2$) [98] were distributed uniformly along the dihedral RC. In all simulations, the terminal base pairs of the duplex were harmonically restrained using positional restraints with a force constant of 5 $kcal/mol/\text{Å}^2$, and in the protein-free bent DNA simulations the backbone phosphate atoms of DNA residues 6, 7, 8, 9 and 20 were additionally restrained with a force constant of 5 $kcal/mol/\text{Å}^2$ similar to the protein-DNA backbone interactions that happen in the crystal structure 1XC8. The other parameters of these umbrella simulations were the same as those in the standard MD simulations sec:MDmethod. The value of the dihedral angle θ was recorded every 0.2 ps. The PMF was calculated from the recorded θ angle using the Weighted Histogram Analysis Method (WHAM) [153] as implemented by Grossfield [151].

4.3 Results and discussions

Although base flipping is generally postulated to occur via the major groove rather than minor groove, the structural analysis of the Fpg-DNA complex shows that the blocking of the major groove by the C-terminal domain of the enzyme and the deformed DNA itself disallows flipping through major groove (see figure 1.12). Thus only the minor groove flipping pathway was concerned. Calculations of the PMF along the θ dihedral angle from 37 independent simulations of each system through the WHAM program result in a free-energy profile (FEP). Histograms of the probability distribution is illustrated in figure 4.8 showing a good convergence of each window and the overlap between one with its neighbouring windows. FEPs for both damaged and undamaged base flipping for the three conditions of the DNA are demonstrated in figure 4.9. The WC base-paired state and the fully flipped-out state of damaged and undamaged DNA for each condition are shown in figure 4.10 to 4.12.

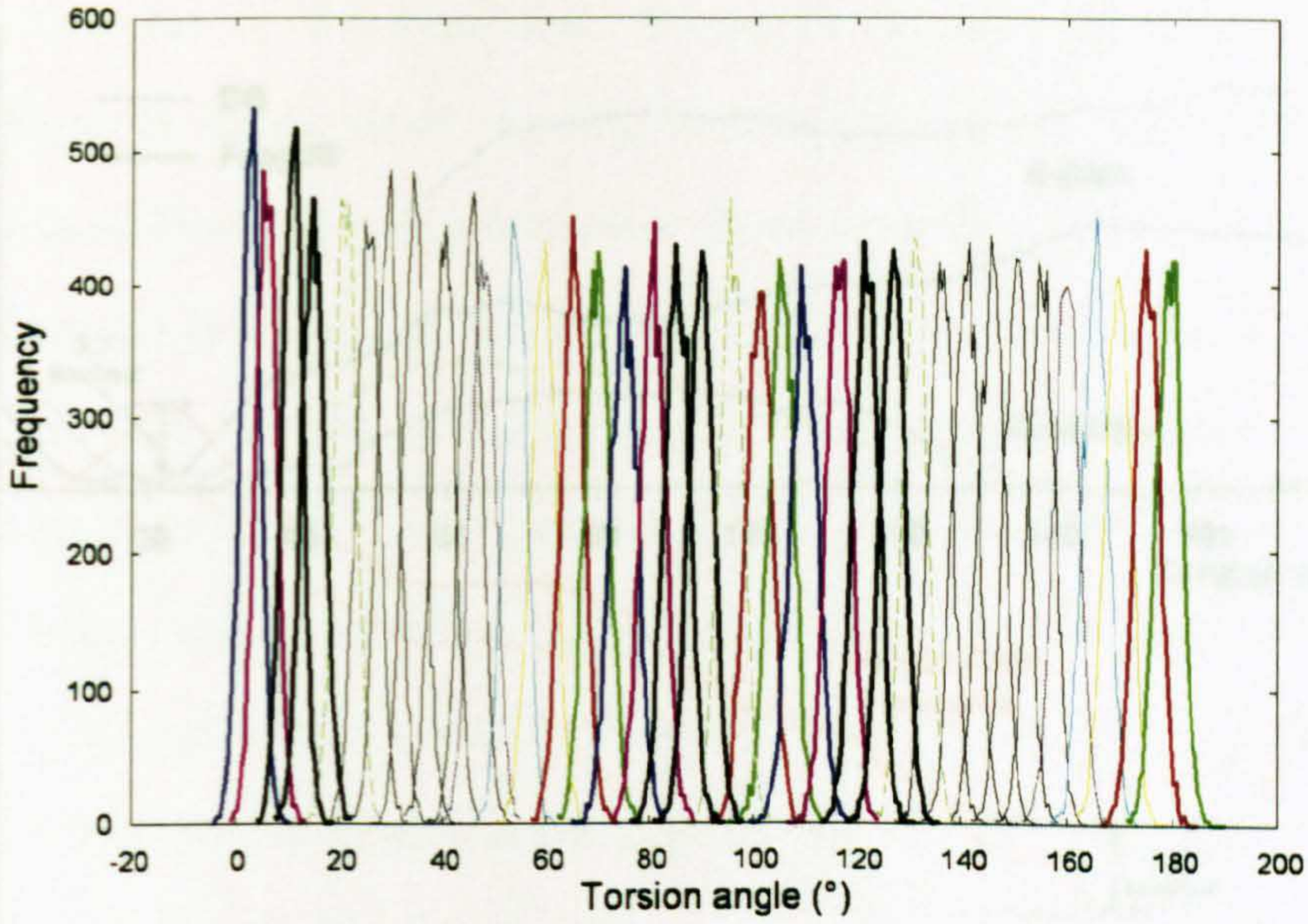


Figure 4.8: Histograms of the probability distribution of each simulation window from 0° to 180° dihedral angle θ.

For G flipping in normal *B*-DNA, rotation of the θ angle from the WC hydrogen bonds (20°) to 40° results in a transient breathing of the orphan C into major groove. Two WC hydrogen bonds are ruptured at 50° rotation and after 70° opening the third hydrogen bond is broken at the cost of 14 $kcal/mol$, which is comparable to the spontaneous base flipping previously reported at 12 $kcal/mol$ [165]. The opening G residue then turns out-of-plane to establish hydrogen bonds with 5'-neighbouring bases and π-π interactions in the T-shaped geometry between two aromatic rings [166]. Opening of FapydG from the helical stack of *B*-DNA slightly differs from G flipping as the FapydG:C base pair is less stable than G:C. The opening of FapydG occurs more easily with the a lower energy barrier of 7 $kcal/mol$.

Base flipping for the distorted DNA occurs in a similar manner as in the normal duplex. However smaller energetic penalties are needed. The G:C base pair is broken at a cost of 5 $kcal/mol$ at 55° rotation. In FapydG flipping, the flat energy around 20° to 35° rotation is due to the stabilisation of FapydG by sliding under its paired cytosine. Opening of 55° ruptures the three hydrogen bonds at a

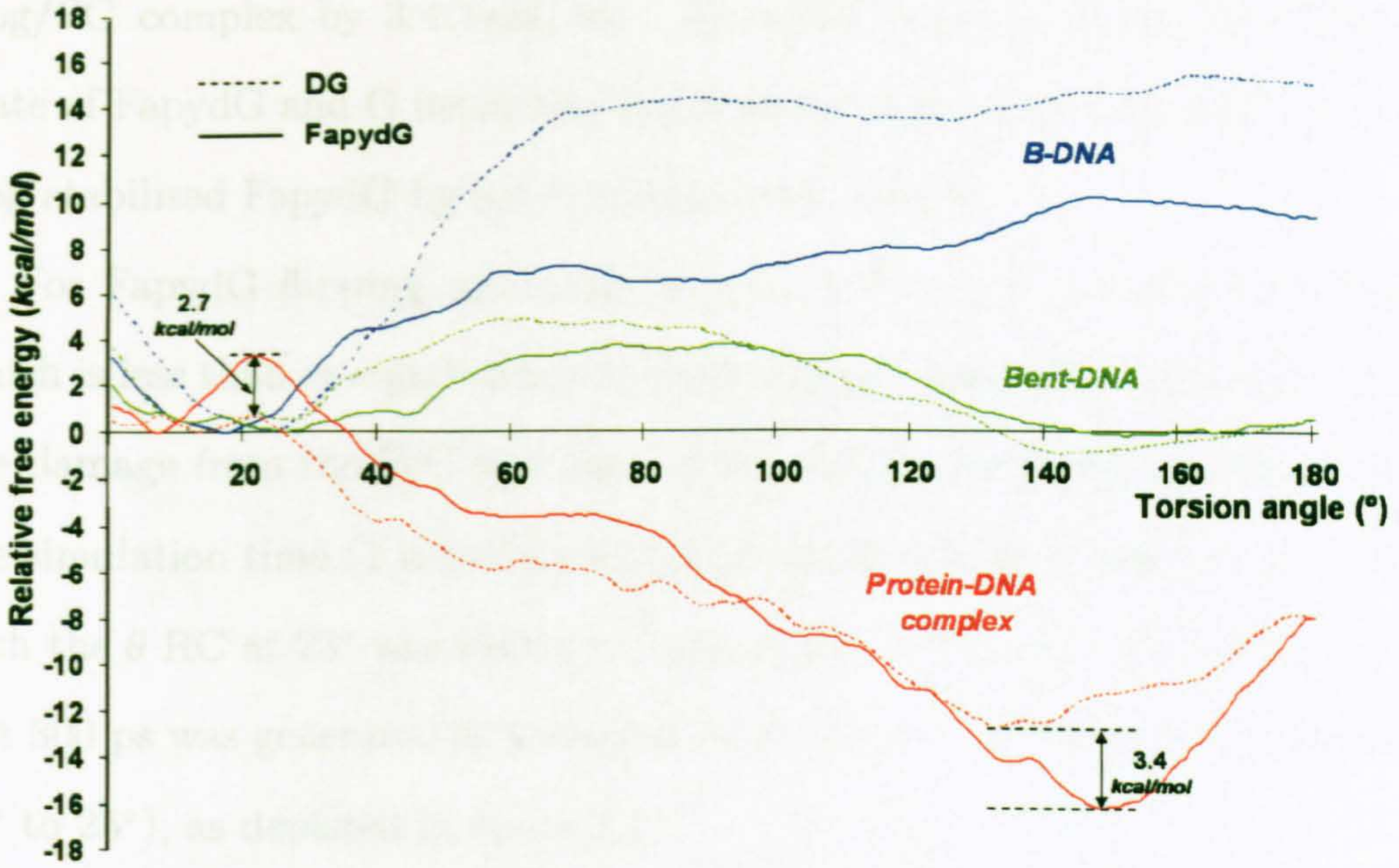


Figure 4.9: Free-energy profiles for minor groove base flipping for *B*-DNA (blue) and bent DNA (green) in aqueous solution, and the Fpg-DNA complex (red). The broken line indicates for the G residue while FapydG is shown by the solid line. Each line was normalised to zero value with respect to its WC base-paired state.

cost of 4 *kcal/mol* and the base turns out-of-plane. It should be noted here that extrahelical G and FapydG are dramatically stabilised by the deformation of the DNA backbone.

In advance, it was expected that the FEP of G flipping into the DNA-binding site of Fpg would reveal a large energy to rotate out of the helix. However, in reality the opening movement of guanine from the stable WC base pair (15°) involves a relatively low energetic penalties of 0.7 *kcal/mol*. After the disruption of G:C hydrogen bonding by the replacement of Arg109 for the target G at 25° rotation, the flipping G residue promptly shifts to an out-of-plane position leading to a large void intrahelically. Two intercalating residues, Met75 and Phe111, subsequently appear to fill the space via minor groove and expel the G residue from the helix by the steric effect. The intercalating residues inhibit the reinsertion of the base. This could suggest that the G flipping pathway is a spontaneously rapid process after the substitution of Arg109. The fully flipped-

out state of G is stabilised by the protein, however it is still less stable than Fpg/FG complex by 3.4 *kcal/mol*. Energetic analysis of the fully flipped-out state of FapydG and G inside the Fpg binding pocket (see table 3.2) showed that Fpg stabilised FapydG by 8.4 *kcal/mol* more than G.

For FapydG flipping, an energy barrier is found at a cost of 3.4 *kcal/mol*, which is less than or equal to one hydrogen bond energy (5-6 *kcal/mol*), to rotate the damage from the WC base pair at 10° to 20°. Recorded torsion angles over the simulation time (1 ns) of an independent simulation of the Fpg/FG complex with the θ RC at 23° was shown in figure 4.13. A time-averaged structure of the last 300 ps was generated as a sample structure over the high energy region (θ = 20° to 25°), as depicted in figure 4.14.

The time-averaged structure of the Fpg/FG complex, $\theta = 21°$, demonstrates the FapydG:C base pair in a buckle conformation while Arg109 is forming a hydrogen bond to the carbonyl group of the orphan C. Interestingly, a water-mediated interaction between Arg260 and the formamide group of FapydG is additionally established. There is no such possible interaction between G and Arg260 during the flipping mechanism. The energy penalty may be required to specifically break this water network to allow the flipping to proceed. Hence it is hypothesised that this indirect interaction between Arg260 and FapydG could direct the discrimination between lesion and non-lesion substrates.

The Arg260 residue, as well as the intercalating triad (Met75, Arg109 and Phe111), are highly conserved in the Fpg superfamily (see figure 3.1). Moreover, Arg260 is a part of a zinc finger motif that functions as a DNA binding domain for the Fpg protein. Mutations of cysteine residues of the zinc finger of the *Ec*Fpg resulted in the lack of DNA binding or cleavage activity [167]. Mutations of the intercalating loop and the conserved residues in the zinc finger of the Endo VIII enzyme revealed that the repair function of the enzyme is impaired, particularly the perturbation at the zinc finger location [168]. However, the precise roles of Arg260 in damage recognition require further studies.

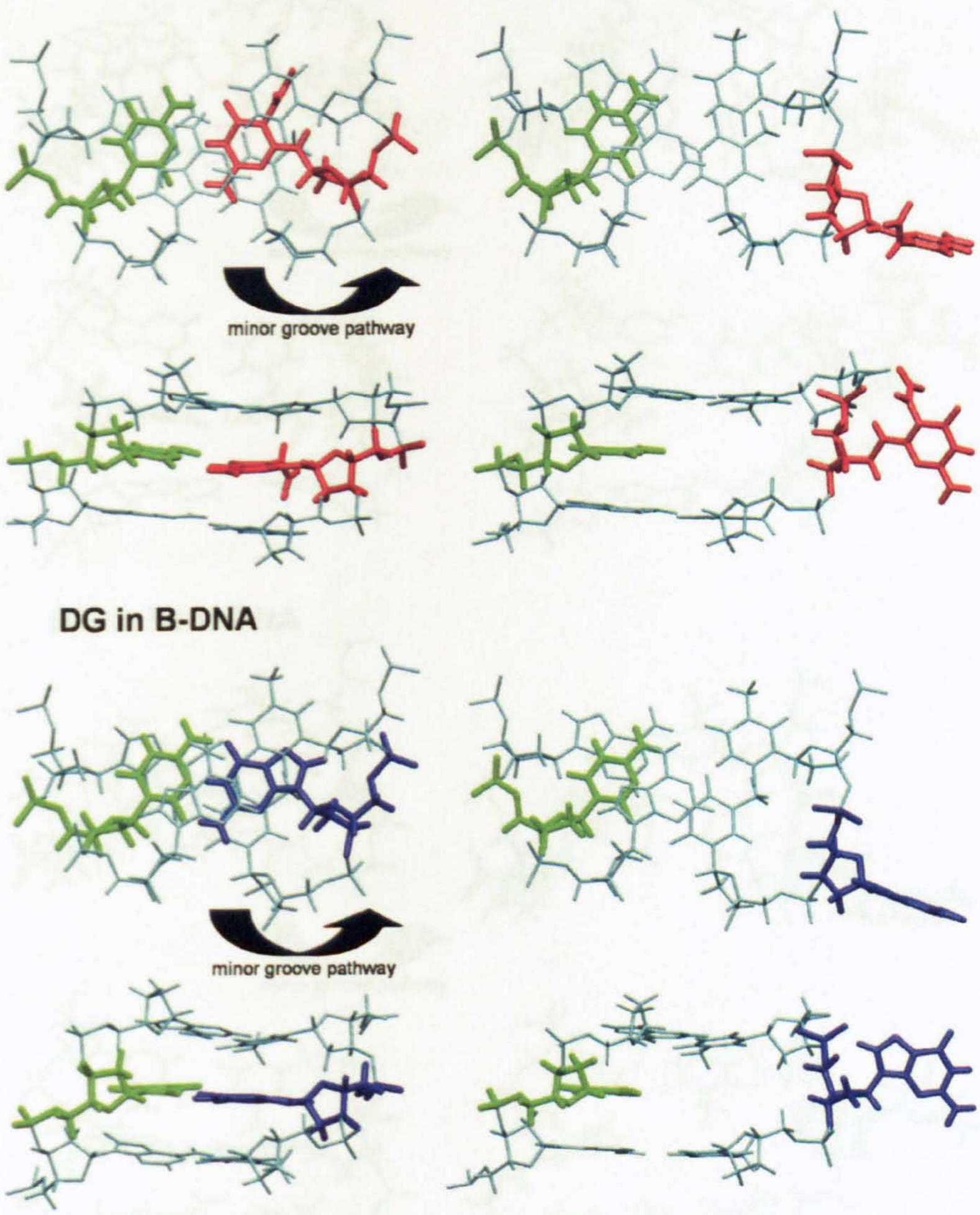


Figure 4.10: The central *B*-DNA tribase at the WC base-paired state in the left panel and the fully flipped-out state in the right panel. FapydG is represented in red, G in blue, C in green and the neighbouring base pairs in cyan. Each system is indicated by top view and minor groove view in top and lower panel respectively.

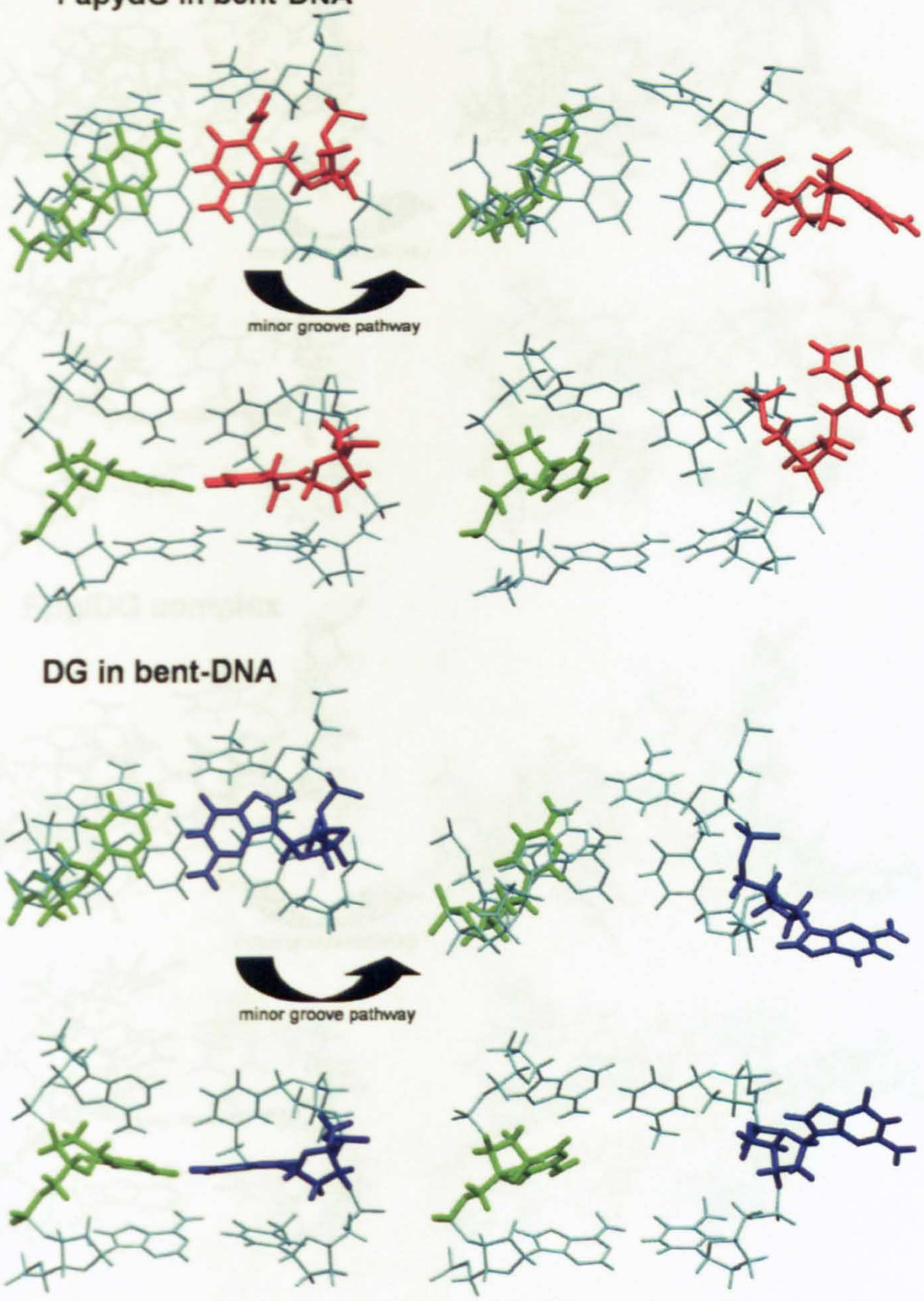


Figure 4.11: The central bent DNA tribase at the WC base-paired state in the left panel and the fully flipped-out state in the right panel. FapydG is represented in red, G in blue, C in green and the neighbouring base pairs in cyan. Each system is indicated by top view and minor groove view in top and lower panel respectively.

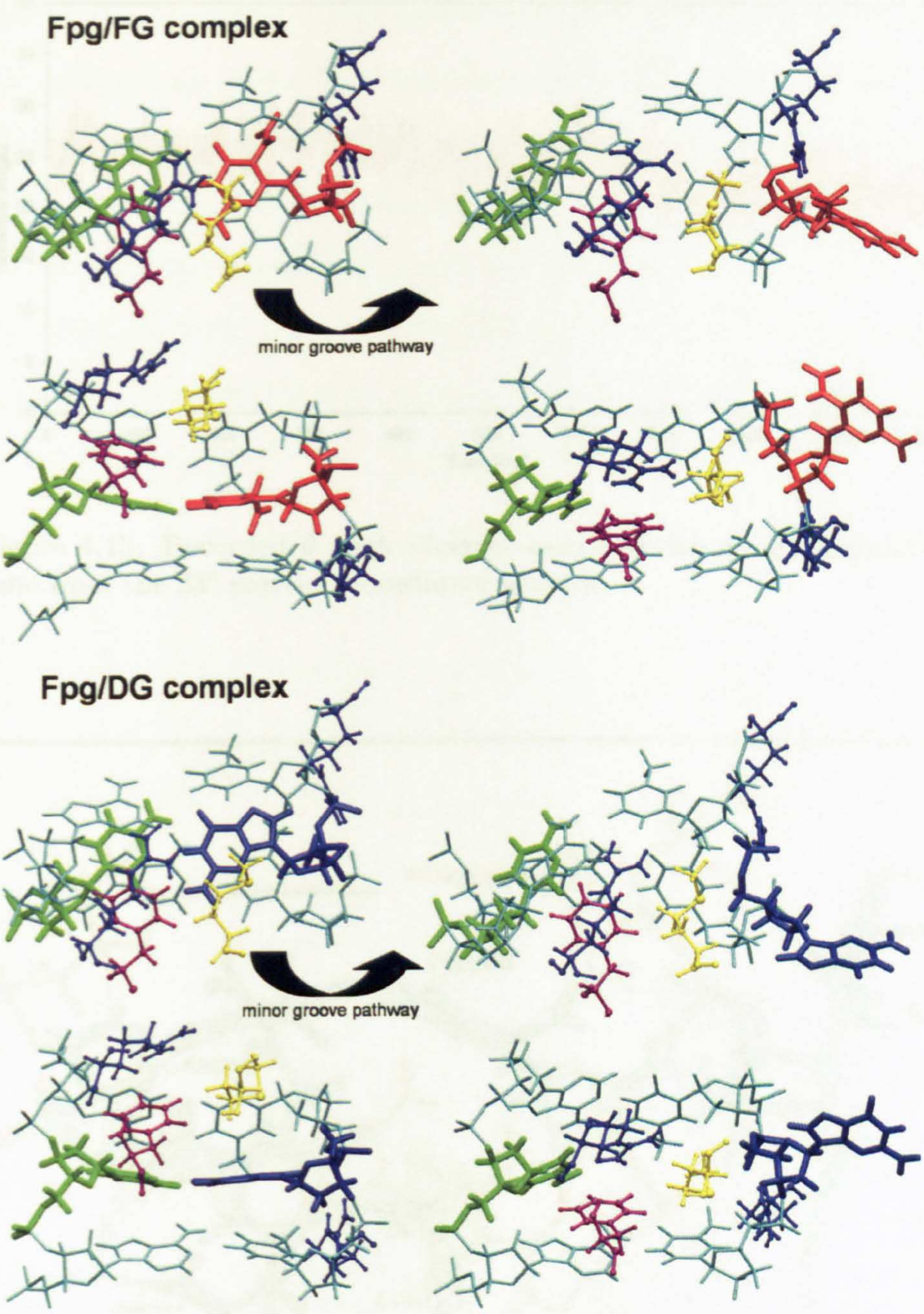


Figure 4.12: The central bent DNA tribase in complexed with Fpg at the WC base-paired state in the left panel and the fully flipped-out state in the right panel. DNA is presented by a licorice model while amino acid residues indicated by the small ball and stick. FapydG is represented in red, G in blue, C in green, the neighbouring base pairs in cyan, Arg109/Arg260 in blue, Met75 in yellow and Phe111 in purple. Each system is indicated by top view and minor groove view in top and lower panel respectively.

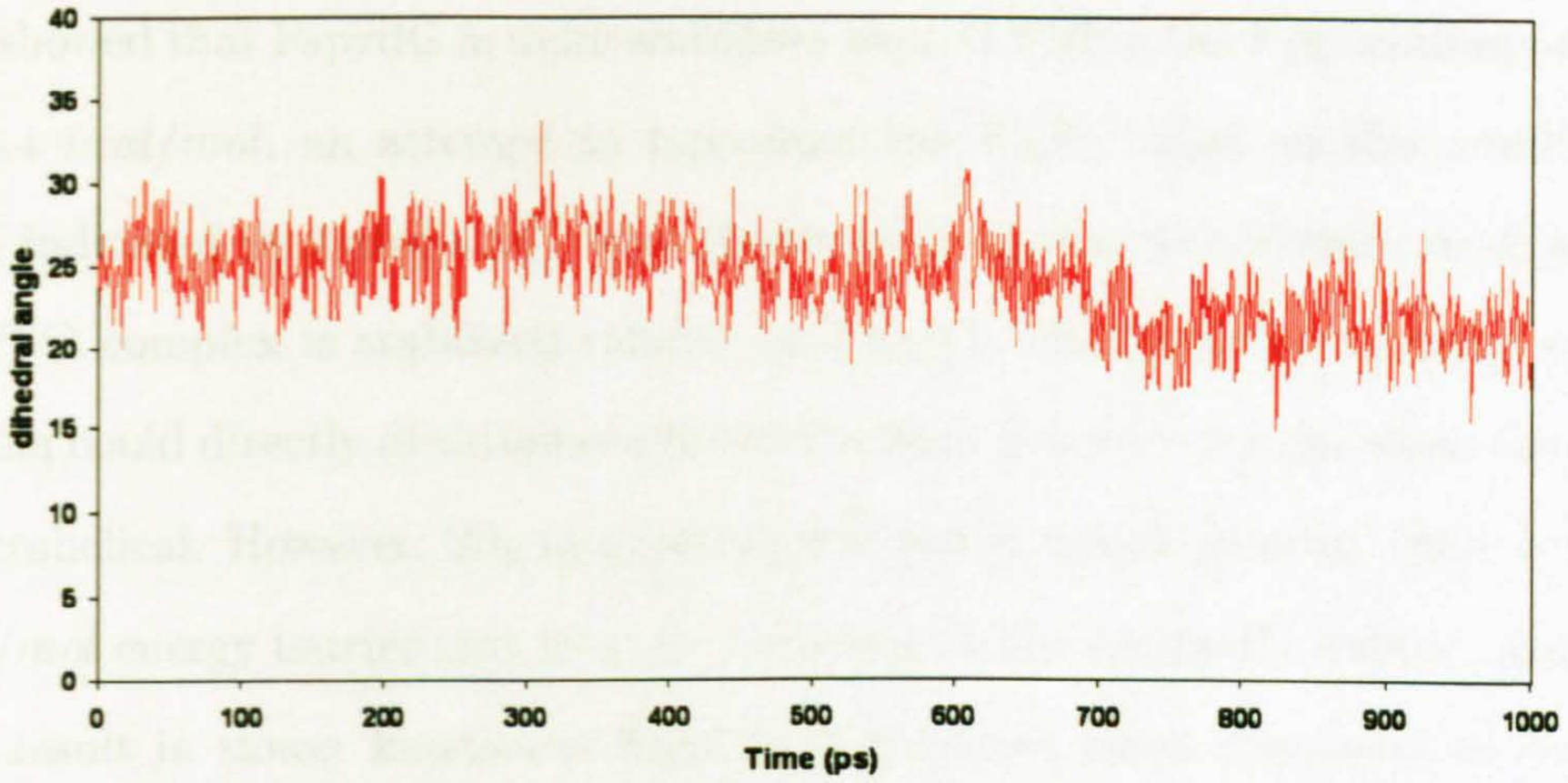


Figure 4.13: Recorded θ angle changes as a function of the simulation time from the 23° reaction coordinate window.

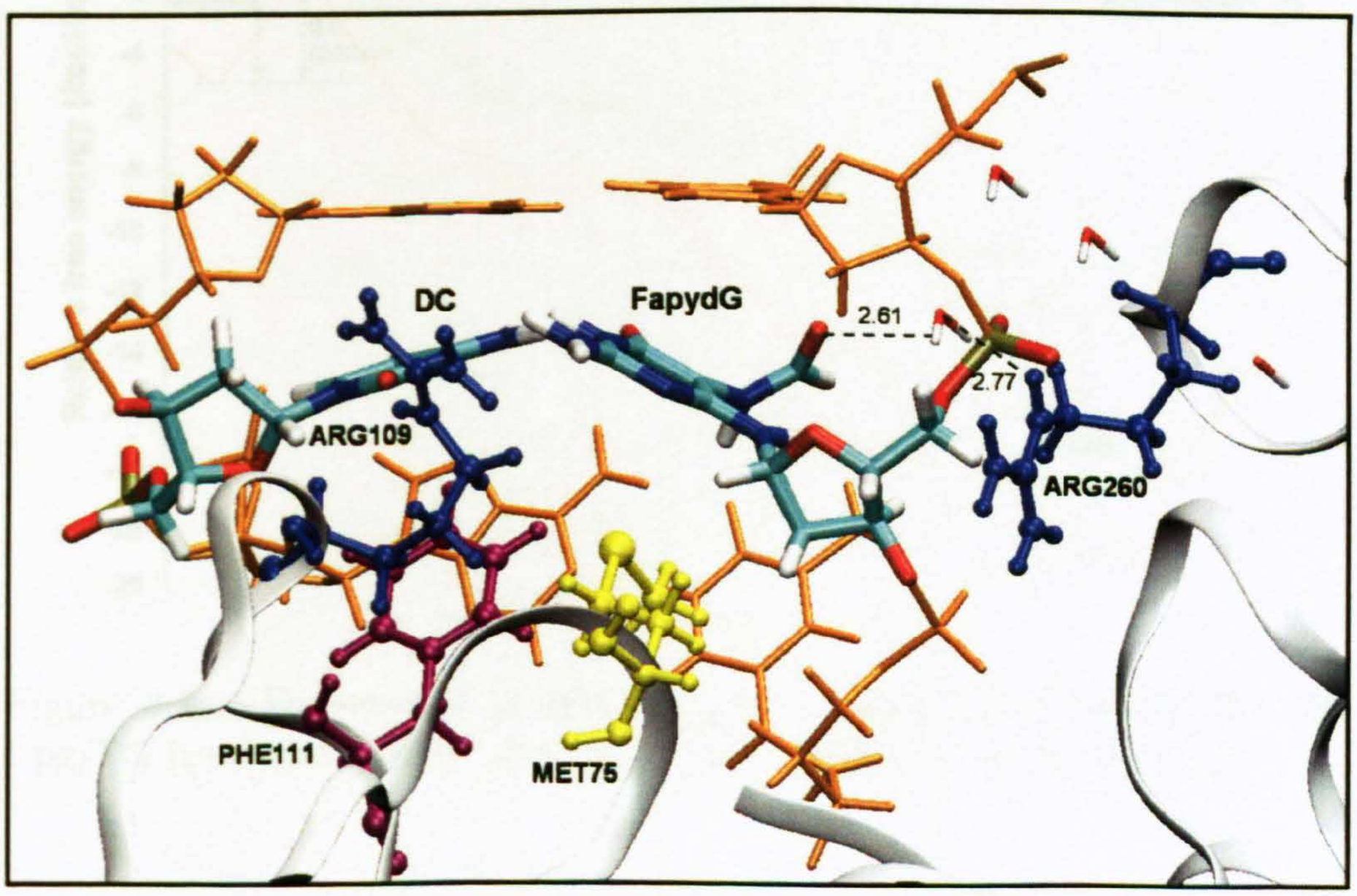


Figure 4.14: Time-averaged structure of the Fpg/FG complex over 300 ps of the 23° dihedral angle window showing a water-mediated interaction between Arg260 and FapydG. DNA is presented by a licorice model while amino acid residues indicated by the small ball and stick. FapydG and C are coloured by atom name, the neighbouring base pairs in orange, Arg109/Arg260 in blue, Met75 in yellow and Phe111 in purple.

Since the MM/GBSA analysis of the Fpg/FG and Fpg/G complexes (table 3.2) showed that FapydG is more stabilised than G within the Fpg binding pocket by 8.4 *kcal/mol*, an attempt to reproduce the FEPs based on this result has been indicated in figure 4.15. It is clearly evident that the initial (pre-flipped) Fpg/FG complex is stabilised relative to Fpg/G. This suggests that the repair protein could directly discriminate the lesion from non-lesions even when the base is intrahelical. However, this recognition now comes with a penalty; there is a 3.4 *kcal/mol* energy barrier that must be overcome to flip a FapydG residue, and this may result in slower kinetics of flipping of damaged bases compared to normal ones.

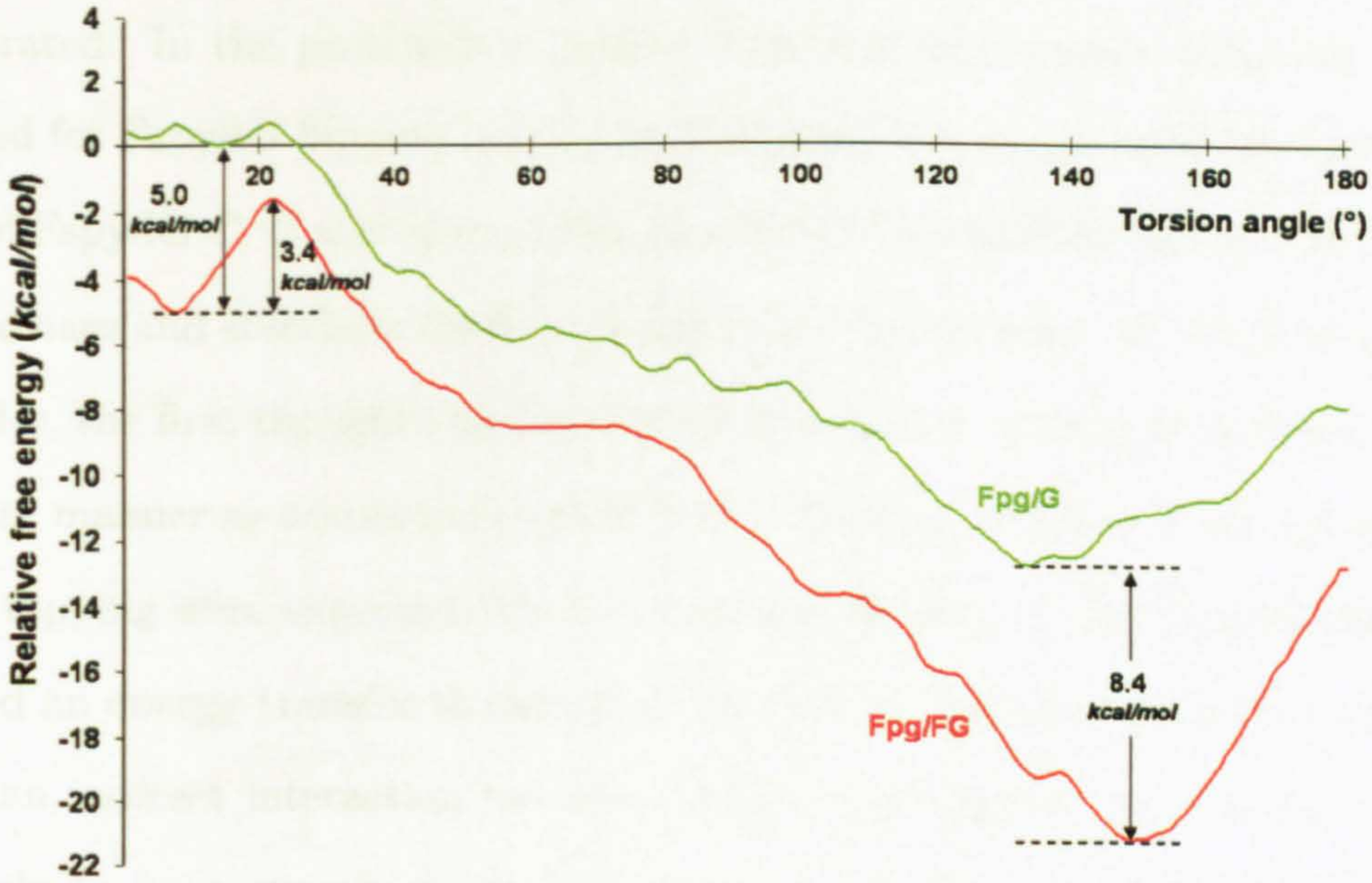


Figure 4.15: Free-energy profiles for minor groove base flipping for the Fpg/FG (red) and Fpg/G (green) complexes based on their $\Delta G_{binding}$.

4.4 Conclusions

Preparation of the appropriate "pre-flipped" models for the Fpg/FG and Fpg/G complexes was an initial task. The "pre-flipped" model was used as a target structure for the targeted MD program to generate a reasonable set of coordi-

nates along the flipping pathway. Obtaining starting structures for each umbrella simulation, the simulations were able to run in parallel processing.

Umbrella sampling in conjunction with the WHAM approach was employed to produce free-energy changes as a function of the base opening dihedral angle θ, C6-C4'-C4'-N3 (FapydG) and C4-C4'-C4'-N3 (DG), into the minor groove. An obvious advantage of using the umbrella sampling method is to improve the sampling of intermediate configurations in high energy area. Finally, FEPs were calculated from the unbiased probability distribution of the chosen reaction coordinate in each sampling window.

Free energy pathways of minor groove flipping for the damaged and undamaged base in *B*-DNA, the distorted DNA, and the protein-bound complex were generated. In the protein-free profiles, only half of energetic penalties were required for FapydG flipping relative to G flipping due to the weak hydrogen bonding of FapydG:C. It was also evident that the DNA bending enhances the opening of the base and stabilises the flipped-out base conformation. In the protein-bound profiles, the first thought was that the protein would facilitate the flipping in the similar manner as occurs in the protein-free profiles. The higher energy penalties of G flipping were expected. On the contrary, flipping of the FapydG residue required an energy transfer to overcome the barrier. Structural analysis suggested that an indirect interaction between Arg260 and FapydG via a water molecule is likely to be a specific recognition between the protein and FapydG whereas Arg109 functions as a recognition probe to rupture the WC hydrogen bonds of the flipping base pair following by the filling the intrahelical gap with the hydrophobic Met75 and Phe111 residues.

The proposed flipping mechanism above supports the hypothesis that damage recognition is dependent on the kinetic rate. Guanine which is a non-lesion substrate can be extruded into the recognition pocket forming a non-specific complex and be released from the pocket in a fast kinetic rate, whereas FapydG is rotated and bound to the Fpg pocket in a good time prior to being cleaved.

Chapter 5

Conclusions & Future Works

Molecular modelling and dynamics simulation techniques were employed for this study in an attempt to understand at the atomistic level, a DNA damage recognition process that is inaccessible through experimental studies. A particular DNA lesion, FapydG, was selected as a case study due to its chemical instability and the difficulty to insert FapydG at a specific site of DNA. Hence theoretical studies are a practical alternative to approach this issue.

5.1 DNA flexibility and damage recognition

Incorporating a modelled FapydG residue into six dodecamer sequences have destabilised the duplex relative to its normal counterpart by 5-10 *kcal/mol*. Although the FapydG:C base pair possesses a hydrogen bonding pattern similar to a WC G:C base pair, destabilisation of FapydG:C is due to the rotatable glycosidic bond of FapydG that disturbs hydrogen bond formation. Dynamic behaviour of DNA in the presence of FapydG was perturbed, and particularly the intrinsic DNA bending property. Damage to DNA appears to enhance the deformation of DNA towards the major groove as being observed in the protein-bound conformation. The PCA indicated that the required bending mode is a major principal component of the dynamics of FapydG-containing DNA. Energetic-related analysis by a novel combined method of the PCA and the D_M was also successfully

introduced as a way to distort the structure from the average configuration to any configuration. This approach indicated that the duplexes are more deformable by protein when FapydG is present. In conclusion, DNA flexibility could act as a 'dynamic signature' used by Fpg to discriminate and locate damaged from undamaged DNA.

To be noted, the unusual dynamic behaviour of 5'-TFT that tends not to deform in the required direction could hinder the recognition of such lesions by Fpg. It was hypothesised that this is due to an unfavourable interaction between the formamide group of FapydG and the 5-methyl group of 3'-T. Further studies could be carried on to verify the issue by substitution of 3'-T with some other nucleobases with and without 5'methyl position such as 5-methylcytosine and uracil, respectively.

5.2 Protein-DNA recognition

Based on the published crystallographic structure of wild-type *Ll*Fpg-DNA containing *c*FapydG complex and other related-structures of the Fpg protein, molecular modelling was employed to construct a model of an extrahelical guanine buried in the binding pocket of Fpg and further pre-flipped models of damaged and undamaged complexes. MD simulations have shown that the FapydG lesion was tightly bound inside the binding pocket with a higher binding free energy of 8.4 *kcal/mol* relative to G. Destabilisation of G inside the binding pocket associated with the fluctuating motion of the undamaged DNA and the αF-β9 loop was noticed. It suggested that the undamaged DNA is loosely bound to Fpg forming a non-specific protein-DNA complex. Interestingly, damage specificity of the Fpg enzyme may depend upon the dynamics of the αF-β9 loop where Arg220 specifically interacts with the formamide functional group of FapydG. It was clearly evident that Fpg is capable of discriminating the lesion from the non-lesion once the base is extrahelical if each base must be presented into the recognition pocket for the damage recognition.

The issue of base flipping was studied using the umbrella sampling method to generate sets of the coordinates in the high energy regions between the flipped-out state and the helical-stacked state; the WHAM approach was then employed to calculate the energetic changes associated with the movement of FapydG and G from inside the duplex to an extrahelical position. Free-energy profiles of damaged and undamaged DNA in *B*-DNA, the distorted DNA, and the protein-bound complex were generated. Base flipping is postulated to occur through minor groove rather than the usual major groove due to the blocking of the major groove by the C-terminal domain of the enzyme and the deformed DNA itself. Thus only minor groove flipping was simulated in this study.

Base-flipping mechanisms of the G base in *B*-DNA and in the distorted DNA exhibited a similar manner as reported elsewhere [165] while FapydG flipping happened with a drastic reduction of energy barrier due to the weak hydrogen bonding of FapydG:C base pair. From the free-energy profiles, it was notable that, in the deformed DNA event, energetic penalties of base flipping were dramatically decreased and the extrahelical base was also stabilised outside the duplex. For the Fpg-DNA system, it was expected that the damaged base flipping would be energetically more favourable. In fact, the calculation showed that FapydG flipping is an unfavourable process compared to G flipping by a little energy barrier of 2.7 $kcal/mol$. This may suggest that the repair protein is capable to specify the lesion with some specific contacts. From the structural analysis, after disruption of FapydG:C hydrogen bonding by the substitution of Arg109, a water-mediated interaction between Arg260 and the formamide group of FapydG is formed before the flipping occurs, whereas there is no such specific interaction between the protein and G. This indicates that G flipping is a spontaneously rapid process. Thus the repair protein may distinguish the lesion from the normal base through the kinetic rate of base flipping. Further experiments could be performed to validate the hypothesis by the mutation of Arg260 in order to reduce the energy barrier and accelerate the flipping process or the damage repair may be halted

due to lack of a specific complex. Alternatively, mutations of the intercalating triad (Arg109, Met75 and Phe111) may suspend or cease base flipping since the disruption of FapydG:C is cancelled. It may be useful for crystallographic experiments to be able to capture the intrahelical state of damaged or undamaged base in the presence of Fpg.

To conclude, an atomistic view of damage recognition by Fpg and the base flipping mechanisms have been revealed using MD simulations. In the case of FapydG, specific recognition is established between the protein and the formamide group and the rotatable glycosidic bond causes FapydG different from its normal counterpart G. Finally, these studies could unravel a comprehensive picture of the repair protein to search and recognise the lesion through the different kinetic rates in which the more deformable damaged DNA is initially located by the protein: the protein subsequently compresses the duplex into an appropriate angle and direction to form a specific protein-DNA complex prior to being flipped and repaired (see figure 5.1). This finding is in good agreement with a lesion-searching model in which the enzyme must form a specific complex with the lesion with a good time to flip and excise the damage from DNA [75].

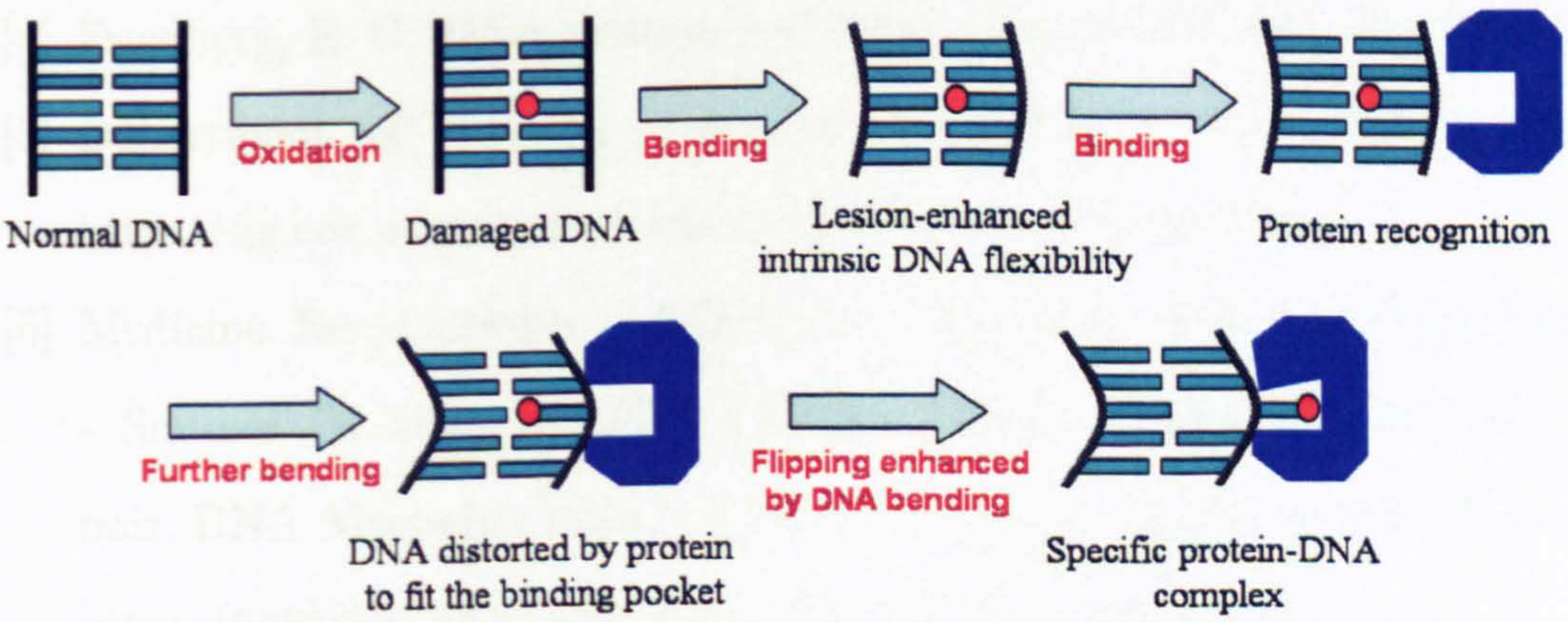


Figure 5.1: A proposed mechanism of DNA damage recognition enhanced by intrinsic DNA curvature.

References

[1] Madden, C. DNA Cartoon. http://www.chrismadden.co.uk/genetics/dna-music.html.

[2] Avery, O. T., MacLeod, C. M., McCarty, M. *Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III.* Journal of Experimental Medicine 79(2):137–158, 1944.

[3] Lodish, H., Berk, A., Matsudaira, P., Kaiser, C. A., Krieger, M., Scott, M. P., Zipursky, S. L., Darnell, J. Molecular biology of the cell. 5 Ed. New York, USA: WH Freeman. 2004: 963.

[4] Friedberg, E. C. *DNA damage and repair.* Nature 421:436–440, 2003.

[5] Department of Biology, University of Miami. Molecular Genetics. http://fig.cox.miami.edu/~cmallery/150/gene/mol_gen.htm.

[6] Medicine Encyclopedia : Genetic in Medicine - Vol. 1. DNA Repair - Sources Of Damage, Base Excision Repair, Nucleotide Excision Repair, DNA Mismatch Repair, Future Directions - Types of DNA Damage. http://medicine.jrank.org/pages/2169/DNA-Repair.html.

[7] Lindahl, T. *Repair of intrinsic DNA lesions.* Mutation Research 238(3):305–311, 1990.

[8] Loeb, L. A. *Apurinic sites as mutagenic intermediates.* Cell 40(3):483–484, 1985.

[9] Pegg, A. E., Dolan, M. E., Moschel, R. C. *Structure, function, and inhi-*

bition of O^6-alkylguanine-DNA alkyltransferase. Progress in Nucleic Acid Research and Molecular Biology 51:167–223, 1995.

[10] Tadokoro, T., Kobayashi, N., Zmudzka, B. Z., Ito, S., Wakamatsu, K., Yamaguchi, Y., Korossy, K. S., Miller, S. A., Beer, J. Z., Hearing, V. J. *UV-induced DNA damage and melanin content in human skin differing in racial/ethnic origin.* FASEB Journal 17(9):1177–1179, 2003.

[11] Marnett, L. J. *Oxyradicals and DNA damage.* Carcinogenesis 21(3):361–370, 2000.

[12] Burrows, C. J., Muller, J. G. *Oxidative nucleobase modifications leading to strand scission.* Chemical Reviews 98(3):1109–1152, 1998.

[13] Steenken, S., Jovanovic, S. V. *How easily oxidizable is DNA? One-electron reduction potentials of adenosine and guanosine radicals in aqueous solution.* Journal of the American Chemical Society 119(3):617–618, 1997.

[14] Beckman, K. B., Ames, B. N. *Oxidative decay of DNA.* Journal of Biological Chemistry 272(32):19633–19636, 1997.

[15] Gilchrest, B. A., Bohr, V. A. *Aging processes, DNA damage, and repair.* FASEB Journal 11(5):322–330, 1997.

[16] Cooke, M. S., Evans, M. D., Dizdaroglu, M., Lunec, J. *Oxidative DNA damage: mechanisms, mutation, and disease.* FASEB Journal 17(10):1195–1214, 2003.

[17] Evans, M. D., Dizdaroglu, M., Cooke, M. S. *Oxidative DNA damage and disease: induction, repair and significance.* Mutation Research 567(1):1–61, 2004.

[18] Malins, D. C., Haimanot, R. *Major alterations in the nucleotide structure of DNA in cancer of the female breast.* Cancer Research 51(19):5430–5432, 1991.

[19] Shimoda, R., Nagashima, M., Sakamoto, M., Yamaguchi, N., Hirohashi, S., Yokota, J., Kasai, H. *Increased formation of oxidative DNA damage, 8-hydroxydeoxyguanosine, in human livers with chronic hepatitis.* Cancer

Research 54(12):3171–3172, 1994.

[20] Crick, F. *The double helix: a personal view.* Nature 248:766–769, 1974.

[21] Friedberg, E. C., Walker, G. C., Siede, W., Wood, R. D., Schultz, R. A., Ellenberger, T. DNA repair and mutagenesis. 2 Ed. Washington, USA: ASM Press. 2006: 5.

[22] Wilson III, D. M., Bohr, V. A. *The mechanics of base excision repair, and its relationship to aging and disease.* DNA Repair 6(4):544–559, 2007.

[23] Prasad, R., Beard, W. A., Strauss, P. R., Wilson, S. H. *Human DNA polymerase beta deoxyribose phosphate lyase. Substrate specificity and catalytic mechanism.* Journal of Biological Chemistry 273(24):15263–15270, 1998.

[24] Dizdaroglu, M. *Base-excision repair of oxidative DNA damage by DNA glycosylases.* Mutation Research 591(1-2):45–59, 2005.

[25] Christmann, M., Tomicic, M. T., Roos, W. P., Kaina, B. *Mechanisms of human DNA repair: an update.* Toxicology 193(1-2):3–34, 2003.

[26] Langie, S. A. S., Knaapen, A. M., Houben, J. M. J., van Kempen, F. C., de Hoon, J. P. J., Gottschalk, R. W. H., Godschalk, R. W. L., van Schooten, F. J. *The role of glutathione in the regulation of nucleotide excision repair during oxidative stress.* Toxicology Letters 168(3):302–309, 2007.

[27] Evans, E., Fellows, J., Coffer, A., Wood, R. D. *Open complex formation around a lesion during nucleotide excision repair provides a structure for cleavage by human XPG protein.* EMBO Journal 16(3):625–638, 1997.

[28] Iyer, R. R., Pluciennik, A., Burdett, V., Modrich, P. L. *DNA mismatch repair: functions and mechanisms.* Chemical Reviews 106(2):302–323, 2006.

[29] Fang, W. H., Modrich, P. *Human strand-specific mismatch repair occurs by a bidirectional mechanism similar to that of the bacterial reaction.* Journal of Biological Chemistry 268(16):11838–11844, 1993.

[30] Thomas, D. C., Roberts, J. D., Kunkel, T. A. *Heteroduplex repair in extracts of human HeLa cells.* Journal of Biological Chemistry 266(6):3744–3751, 1991.

[31] Yang, W. *Structure and function of mismatch repair proteins.* Mutation Research 460(3-4):245–256, 2000.

[32] Daniels, D. S., Woo, T. T., Luu, K. X., Noll, D. M., Clarke, N. D., Pegg, A. E., Tainer, J. A. *DNA binding and nucleotide flipping by the human DNA repair protein AGT.* Nature 11(8):714–720, 2004.

[33] Trewick, S. C., Henshaw, T. F., Hausinger, R. P., Lindahl, T., Sedgwick, B. *Oxidative demethylation by Escherichia coli AlkB directly reverts DNA base damage.* Nature 419:174–178, 2002.

[34] Falnes, P. O., Johansen, R. F., Seeberg, E. *AlkB-mediated oxidative demethylation reverses DNA damage in Escherichia coli.* Nature 419:178–182, 2002.

[35] Helleday, T. *Pathways for mitotic homologous recombination in mammalian cells.* Mutation Research 532(1-2):103–115, 2003.

[36] Nick McElhinny, S. A., Snowden, C. M., McCarville, J., Ramsden, D. A. *Ku recruits the XRCC4-ligase IV complex to DNA ends.* Molecular and Cellular Biology 20(9):2996–3003, 2000.

[37] Holliday, R. *A mechanism for gene conversion in fungi.* Genetical Research 5:282–304, 1964.

[38] Lehmann, A. R., Niimi, A., Ogi, T., Brown, S., Sabbioneda, S., Wing, J. F., Kannouche, P. L., Green, C. M. *Translesion synthesis: Y-family polymerases and the polymerase switch.* DNA Repair 6(7):891–899, 2007.

[39] McCulloch, S. D., Kokoska, R. J., Masutani, C., Iwai, S., Hanaoka, F., Kunkel, T. A. *Preferential cis-syn thymine dimer bypass by DNA polymerase η occurs with biased fidelity.* Nature 428:97–100, 2004.

[40] Crespo-Hernandez, C. E., Arce, R. *Formamidopyrimidines as major products in the low- and high-intensity UV irradiation of guanine derivatives.* Journal of Photochemistry and Photobiology B: Biology 73:167–175, 2004.

[41] Douki, T., Martini, R., Ravanat, J. L., Turesky, R. J., Cadet, J. *Measurement of 2,6-diamino-4-hydroxy-5-formamidopyrimidine and 8-oxo-7,8-*

dihydroguanine in isolated DNA exposed to gamma radiation in aqueous solution. Carcinogenesis 18(12):2385–2391, 1997.

[42] Pouget, J. P., Douki, T., Richard, M. J., Cadet, J. *DNA damage induced in cells by gamma and UVA radiation as measured by HPLC/GC-MS and HPLC-EC and Comet assay.* Chemical Research in Toxicology 13(7):541–549, 2000.

[43] Hu, J., de Souza-Pinto, N. C., Haraguchi, K., Hogue, B. A., Jaruga, P., Greenberg, M. M., Dizdaroglu, M., Bohr, V. A. *Repair of formamidopyrimidines in DNA involves different glycosylases: role of the OGG1, NTH1, and NEIL1 enzymes.* Journal of Biological Chemistry 280(49):40544–440551, 2005.

[44] Kasprzak, K. S., Jaruga, P., Zastawny, T. H., North, S. L., Riggs, C. W., Olinski, R., Dizdaroglu, M. *Oxidative DNA base damage and its repair in kidneys and livers of nickel(II)-treated male F344 rats.* Carcinogenesis 18(2):271–277, 1997.

[45] Dizdaroglu, M., Olinski, R., Doroshow, J. H., Akman, S. A. *Modification of DNA bases in chromatin of intact target human cells by activated human polymorphonuclear leukocytes.* Cancer Research 53(6):1269–1272, 1993.

[46] Mori, T., Hori, Y., Dizdaroglu, M. *DNA base damage generated in vivo in hepatic chromatin of mice upon whole body gamma-irradiation.* International Journal of Radiation Biology 64(6):645–650, 1993.

[47] Wiederholt, C. J., Greenberg, M. M. *Fapy·dG instructs Klenow exo(-) to misincorporate deoxyadenosine.* Journal of the American Chemical Society 124(25):7278–7279, 2002.

[48] Tudek, B. *Imidazole ring-opened DNA purines and their biological significance.* Journal of Biochemistry and Molecular Biology 31(1):12–19, 2003.

[49] Patro, J. N., Haraguchi, K., Delaney, M. O., Greenberg, M. M. *Probing the configurations of formamidopyrimidine lesions Fapy·dA and Fapy·dG in DNA using endonuclease IV.* Biochemistry 43(42):13397–13403, 2004.

[50] Haraguchi, K., Greenberg, M. M. *Synthesis of oligonucleotides containing Fapy·dG (N6-(2-deoxy-α,β-D-erythro-pentofuranosyl)-2,6-diamino-4-hydroxy-5-formamidopyrimidine).* Journal of the American Chemical Society 123(35):8636–8637, 2001.

[51] Delaney, M. O., Greenberg, M. M. *Synthesis of oligonucleotides and thermal stability of duplexes containing the β-C-nucleoside analogue of Fapy·dG.* Chemical Research in Toxicology 15(11):1460–1465, 2002.

[52] Wiederholt, C. J., Delaney, M. O., Pope, M. A., David, S. S., Greenberg, M. M. *Repair of DNA containing Fapy·dG and its β-C-nucleoside analogue by formamidopyrimidine DNA glycosylase and MutY.* Biochemistry 42(32):9755–9760, 2003.

[53] Ober, M., Linne, U., Gierlich, J., Carell, T. *The two main DNA lesions 8-oxo-7,8-dihydroguanine and 2,6-diamino-5-formamido-4-hydroxypyrimidine exhibit strongly different pairing properties.* Angewandte Chemie 115(40):5097–5101, 2003.

[54] Coste, F., Ober, M., Carell, T., Boiteux, S., Zelwer, C., Castaing, B. *Structural basis for the recognition of the Fapy·dG lesion (2,6-diamino-4-hydroxy-5-formamidopyrimidine) by formamidopyrimidine-DNA glycosylase.* Journal of Biological Chemistry 279(42):44074–44083, 2004.

[55] The Protein Data Bank. http://www.pdb.org.

[56] Serre, L., de Jsus, K. P., Boiteux, S., Zelwer, C., Castaing, B. *Crystal structure of the Lactococcus lactis formamidopyrimidine-DNA glycosylase bound to an abasic site analogue-containing DNA.* EMBO Journal 21(12):2854–2865, 2002.

[57] Baker, N. A., Sept, D., Joseph, S., Holst, M. J., McCammon, J. A. *Electrostatics of nanosystems: application to microtubules and the ribosome.* Proceedings of the National Academy of Sciences of the United States of America 98(18):10037–10041, 2001.

[58] Huffman, J. L., Sundheim, O., Tainer, J. A. *DNA base damage recognition*

and removal: new twists and grooves. Mutation Research 577(1-2):55–76, 2005.

[59] Pensak, D. A. *Molecular modelling: scientific and technological boundaries.* Pure and Applied Chemistry 61(3):601–603, 1989.

[60] Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., Weiner, P. *A new force field for molecular mechanical simulation of nucleic acids and proteins.* Journal of the American Chemical Society 106(3):765–784, 1984.

[61] van Gunsteren, W. F., Berendsen, H. J. C. *Algorithms for macromolecular dynamics and constraint dynamics.* Molecular Physics 34(5):1311–1327, 1977.

[62] Verlet, L. *Computer experiments on classical fluids. I. Thermodynamical properties of lennard-jones molecules.* Physical Review 159(1):98–103, 1967.

[63] Case, D. A., Darden, T. A., Cheatham III, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Merz, K. M., Wang, B., Pearlman, D. A., Crowley, M., Brozell, S., Tsui, V., Gohlke, H., Mongan, J., Hornak, V., Cui, G., Beroza, P., Schafmeister, C., Caldwell, J. W., Ross, W. S., Kollman, P. A. AMBER 8. 2004.

[64] Case, D. A., Darden, T. A., Cheatham III, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Merz, K. M., Pearlman, D. A., Crowley, M., Walker, R. C., Zhang, W., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Wong, K. F., Paesani, F., Wu, X., Brozell, S., Tsui, V., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Beroza, P., Mathews, D. H., Schafmeister, C., Ross, W. S., Kollman, P. A. AMBER 9. 2006.

[65] Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., Woods, R. J. *The Amber biomolecular simulation programs.* Journal of Computational Chemistry 26(16):1668–1688, 2005.

[66] Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W.,

Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., Kollman, P. *A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations.* Journal of Computational Chemistry 24(16):1999–2012, 2003.

[67] Humphrey, W., Dalke, A., Schulten, K. *VMD – Visual Molecular Dynamics.* Journal of Molecular Graphics 14:33–38, 1996.

[68] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., Ferrin, T. E. *UCSF Chimera–A visualization system for exploratory research and analysis.* Journal of Computational Chemistry 25(13):1605–1612, 2004.

[69] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., Klein, M. L. *Comparison of simple potential functions for simulating liquid water.* Journal of Chemical Physics 79(2):926–935, 1983.

[70] Várnai, P., Zakrzewska, K. *DNA and its counterions: a molecular dynamics study.* Nucleic Acids Research 32(14):4269–4280, 2004.

[71] Shields, G. C., Laughton, C. A., Orozco, M. *Molecular dynamics simulation of a PNA·DNA·PNA triple helix in aqueous solution.* Journal of the American Chemical Society 120(24):5895–5904, 1998.

[72] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., Haak, J. R. *Molecular dynamics with coupling to an external bath.* Journal of Chemical Physics 81(8):3684–3690, 1984.

[73] Ryckaert, J.-P., Ciccotti, G., Berendsen, H. J. C. *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* Journal of Computational Physics 23(3):327–341, 1977.

[74] Darden, T., York, D., Pedersen, L. *Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems.* Journal of Chemical Physics 98(12):10089–10092, 1993.

[75] Blainey, P. C., van Oijen, A. M., Banerjee, A., Verdine, G. L., Xie, X. S.

A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA. Proceedings of the National Academy of Sciences of the United States of America 103(15):5752–5757, 2006.

[76] Hogan, M. E., Austin, R. H. *Importance of DNA stiffness in protein-DNA binding specificity.* Nature 329:263–266, 1987.

[77] Allemann, R. K., Egli, M. *DNA recognition and bending.* Chemistry & Biology 4(9):643–650, 1997.

[78] Pastor, N., Weinstein, H. Protein-DNA interactions in the initiation of transcription: the role of flexibility and dynamics of the TATA recognition sequence and the TATA box binding protein. In: Theoretical Biochemistry - Processes and Properties of Biological Systems. Eriksson, L. A. ed. volume 9 Ed. Elsevier 2001 377–407.

[79] Schwabe, J. W. *The role of water in protein-DNA interactions.* Current Opinion in Structural Biology 7(1):126–134, 1997.

[80] Reddy, C. K., Das, A., Jayaram, B. *Do water molecules mediate protein-DNA recognition?* Journal of Molecular Biology 314(3):619–632, 2001.

[81] Koo, H. S., Wu, H. M., Crothers, D. M. *DNA bending at adenine·thymine tracts.* Nature 320:501–506, 1986.

[82] Goodsell, D. S., Kopka, M. L., Cascio, D., Dickerson, R. E. *Crystal structure of CATGGCCATG and its implications for A-tract bending models.* Proceedings of the National Academy of Sciences of the United States of America 90(7):2930–2934, 1993.

[83] Grzeskowiak, K., Goodsell, D. S., Kaczor-Grzeskowiak, M., Cascio, D., Dickerson, R. E. *Crystallographic analysis of C-C-A-A-G-C-T-T-G-G and its implications for bending in B-DNA.* Biochemistry 32(34):8923–8931, 1993.

[84] Dickerson, R. E., Goodsell, D., Kopka, M. L. *MPD and DNA bending in crystals and in solution.* Journal of Molecular Biology 256(1):108–125, 1996.

[85] Crothers, D. M., Haran, T. E., Nadeau, J. G. *Intrinsically bent DNA.* Journal of Biological Chemistry 265(13):7093–7096, 1990.

[86] Lavery, R., Sklenar, H. CURVES 5.1. Helical analysis of irregular nucleic acids. 1996.

[87] Bruner, S. D., Norman, D. P. G., Verdine, G. L. *Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA.* Nature 403:859–866, 2000.

[88] Hollis, T., Ichikawa, Y., Ellenberger, T. *DNA bending and a flip-out mechanism for base excision by the helix-hairpin-helix DNA glycosylase, Escherichia coli AlkA.* EMBO Journal 19(4):758–766, 2000.

[89] Lau, A. Y., Wyatt, M. D., Glassner, B. J., Samson, L. D., Ellenberger, T. *Molecular basis for discriminating between normal and damaged bases by the human alkyladenine glycosylase, AAG.* Proceedings of the National Academy of Sciences of the United States of America 97(25):13573–13578, 2000.

[90] Parikh, S. S., Mol, C. D., Slupphaug, G., Bharati, S., Krokan, H. E., Tainer, J. A. *Base excision repair initiation revealed by crystal structures and binding kinetics of human uracil-DNA glycosylase with DNA.* EMBO Journal 17(17):5214–5226, 1998.

[91] Malins, D. C., Polissar, N. L., Ostrander, G. K., Vinson, M. A. *Single 8-oxoguanine and 8-oxo-adenine lesions induce marked changes in the backbone structure of a 25-base DNA strand.* Proceedings of the National Academy of Sciences of the United States of America 97(23):12442–12445, 2000.

[92] Miller, J. H., Fan-Chiang, C.-C. P., Straatsma, T. P., Kennedy, M. A. *8-Oxoguanine enhances bending of DNA that favors binding to glycosylases.* Journal of the American Chemical Society 125(20):6331–6336, 2003.

[93] Mao, H., Deng, Z., Wang, F., Harris, T. M., Stone, M. P. *An intercalated and thermally stable FAPY adduct of aflatoxin B1 in a DNA duplex: structural refinement from 1H NMR.* Biochemistry 37(13):4374–4387, 1998.

[94] Frisch, J. M., Trucks, W. G., Schlegel, B. H., Scuseria, E. G., Robb, A. M., Cheeseman, R. J., Zakrzewski, G. V., Montgomery, A. J., Stratmann, E. R., Burant, C. J., Dapprich, S., Millam, M. J., Daniels, D. A., Kudin, N. K., Strain, C. M., Farkas, O., Tomasi, J., Barone, V., Cossi, M., Cammi, R., Mennucci, B., Pomelli, C., Adamo, C., Clifford, S., Ochterski, J., Petersson, A. G., Ayala, Y. P., Cui, Q., Morokuma, K., Malick, K. D., Rabuck, D. A., Raghavachari, K., Foresman, B. J., Cioslowski, J., Ortiz, V. J., Baboul, G. A., Stefanov, B. B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Gomperts, R., Martin, L. R., Fox, J. D., Keith, T., Al-Laham, A. M., Peng, Y. C., Nanayakkara, A., Challacombe, M., Gill, W. M. P., Johnson, B., Chen, W., Wong, W. M., Andres, L. J., Gonzalez, C., Head-Gordon, M., Replogle, S. E., Pople, A. J. Gaussian 98, Revision A.9. 1998.

[95] Bayly, C. I., Cieplak, P., Cornell, W., Kollman, P. A. *A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model.* Journal of Physical Chemistry 97(40):10269–10280, 1993.

[96] Wang, J., Cieplak, P., Kollman, P. A. *How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?* Journal of Computational Chemistry 21(12):1049–1074, 2000.

[97] Harris, S. A., Gavathiotis, E., Searle, M. S., Orozco, M., Laughton, C. A. *Cooperativity in drug-DNA recognition: a molecular dynamics study.* Journal of the American Chemical Society 123(50):12658–12663, 2001.

[98] Beardsell, M. A. *DNA damage recognition and the inhibition of its repair.* PhD thesis. School of Pharmacy, University of Nottingham. 2005.

[99] Amadei, A., Linssen, A. B. M., Berendsen, H. J. C. *Essential dynamics of proteins.* Proteins-Structure Function and Genetics 17(4):412–425, 1993.

[100] Pérez, A., Blas, J. R., Rueda, M., López-Bes, J. M., delaCruz, X., Orozco, M. *Exploring the essential dynamics of B-DNA.* Journal of Chemical The-

ory and Computation 1(5):790–800, 2005.

[101] Wlodek, S. T., Clark, T. W., Scott, L. R., McCammon, J. A. *Molecular dynamics of acetylcholinesterase dimer complexed with Tacrine.* Journal of the American Chemical Society 119(40):9513–9522, 1997.

[102] Sherer, E. C., Harris, S. A., Soliva, R., Orozco, M., Laughton, C. A. *Molecular dynamics studies of DNA A-Tract structure and flexibility.* Journal of the American Chemical Society 121(25):5981–5991, 1999.

[103] Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A., Case, D. A. *Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices.* Journal of the American Chemical Society 120(37):9401–9409, 1998.

[104] Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., Cheatham, T. E. *Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models.* Accounts of Chemical Research 33(12):889–897, 2000.

[105] Sitkoff, D., Sharp, K. A., Honig, B. *Accurate calculation of hydration free energies using macroscopic solvent models.* Journal of Physical Chemistry 98(7):1978–1988, 1994.

[106] Onufriev, A., Bashford, D., Case, D. A. *Modification of the generalized born model suitable for macromolecules.* Journal of Physical Chemistry B 104:3712–3720, 2000.

[107] Kormos, B. L., Beveridge, D. L. Ras-Raf MM_PB(GB)SA Tutorial: AMBER 8. http://amber.scripps.edu/tutorial/AMBER-MM_PBSA-tutorial_v8.pdf.

[108] Still, W. C., Tempczyk, A., Hawley, R. C., Hendrickson, T. *Semianalytical treatment of solvation for molecular mechanics and dynamics.* Journal of the American Chemical Society 112(16):6127–6129, 1990.

[109] Case, D. A., Pearlman, D. A., Caldwell, J. W., Cheatham, T. E., Wang, J.,

Ross, W. S., Simmerling, C. L., Darden, T. A., Merz, K. M., Stanton, R. V., Cheng, A. L., Vincent, J. J., Crowley, M., Tsui, V., Gohlke, H., Radmer, R. J., Duan, Y., Pitera, J., Massova, I., Seibel, G. L., Singh, U. C., Weiner, P. K., Kollman, P. A. AMBER 7. 2002.

[110] Dickerson, R. E. *DNA bending: the prevalence of kinkiness and the virtues of normality.* Nucleic Acids Research 26(8):1906–1926, 1998.

[111] Lavery, R., Sklenar, H. *The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids.* Journal of Biomolecular Structure and Dynamics 6(1):63–91, 1988.

[112] Mahalanobis, P. C. *On the generalised distance in statistics.* In *Proceedings of the National Institute of Science of India* (1936). Vol. 12. .

[113] Meyer, T., Ferrer-Costa, C., Pérez, A., Rueda, M., Bidon-Chanal, A., Luque, F. J., Laughton, C. A., Orozco, M. *Essential dynamics: a tool for efficient trajectory compression and management.* Journal of Chemical Theory and Computation 2(2):251–258, 2006.

[114] Perlow-Poehnelt, R. A., Zharkov, D. O., Grollman, A. P., Broyde, S. *Substrate discrimination by formamidopyrimidine-DNA glycosylase: distinguishing interactions within the active site.* Biochemistry 43(51):16092–16105, 2004.

[115] Ober, M., Mller, H., Pieck, C., Gierlich, J., Carell, T. *Base pairing and replicative processing of the formamidopyrimidine-dG DNA lesion.* Journal of the American Chemical Society 127(51):18143–18149, 2005.

[116] Seibert, E., Ross, J. B. A., Osman, R. *Role of DNA flexibility in sequence-dependent activity of uracil DNA glycosylase.* Biochemistry 41(36):10976–10984, 2002.

[117] Eftedal, I., Guddal, P. H., Slupphaug, G., Volden, G., Krokan, H. E. *Consensus sequences for good and poor removal of uracil from double stranded DNA by uracil-DNA glycosylase.* Nucleic Acids Research 21(9):2095–2101, 1993.

[118] Nilsen, H., Yazdankhah, S. P., Eftedal, I., Krokan, H. E. *Sequence specificity for removal of uracil from U·A pairs and U·G mismatches by uracil-DNA glycosylase from Escherichia coli, and correlation with mutational hotspots.* FEBS Letters 362(2):205–209, 1995.

[119] Slupphaug, G., Eftedal, I., Kavli, B., Bharati, S., Helle, N. M., Haug, T., Levine, D. W., Krokan, H. E. *Properties of a recombinant human uracil-DNA glycosylase from the UNG gene and evidence that UNG encodes the major uracil-DNA glycosylase.* Biochemistry 34(1):128–138, 1995.

[120] Verdine, G. L., Bruner, S. D. *How do DNA repair proteins locate damaged bases in the genome?* Chemistry & Biology 4(5):329–334, 1997.

[121] Sugahara, M., Mikawa, T., Kumasaka, T., Yamamoto, M., Kato, R., Fukuyama, K., Inoue, Y., Kuramitsu, S. *Crystal structure of a repair enzyme of oxidatively damaged DNA, MutM (Fpg), from an extreme thermophile, Thermus thermophilus HB8.* EMBO Journal 19(15):3857–3869, 2000.

[122] Gilboa, R., Zharkov, D. O., Golan, G., Fernandes, A. S., Gerchman, S. E., Matz, E., Kycia, J. H., Grollman, A. P., Shoham, G. *Structure of formamidopyrimidine-DNA glycosylase covalently complexed to DNA.* Journal of Biological Chemistry 277(22):19811–19816, 2002.

[123] Fromme, J. C., Verdine, G. L. *DNA lesion recognition by the bacterial repair enzyme MutM.* Journal of Biological Chemistry 278(51):51543–51548, 2003.

[124] Jiang, D., Hatahet, Z., Blaisdell, J. O., Melamede, R. J., Wallace, S. S. *Escherichia coli endonuclease VIII: cloning, sequencing, and overexpression of the nei structural gene and characterization of nei and nei nth mutants.* Journal of Bacteriology 179(11):3773–3782, 1997.

[125] Jiang, D., Hatahet, Z., Melamede, R. J., Kow, Y. W., Wallace, S. S. *Characterization of Escherichia coli endonuclease VIII.* Journal of Biological Chemistry 272(51):32230–32239, 1997.

[126] Melamede, R. J., Hatahet, Z., Kow, Y. W., Ide, H., Wallace, S. S. *Isolation and characterization of endonuclease VIII from Escherichia coli.* Biochemistry 33(5):1255–1264, 1994.

[127] Purmal, A. A., Lampman, G. W., Bond, J. P., Hatahet, Z., Wallace, S. S. *Enzymatic processing of uracil glycol, a major oxidative product of DNA cytosine.* Journal of Biological Chemistry 273(16):10026–10035, 1998.

[128] Amara, P., Serre, L. *Functional flexibility of Bacillus stearothermophilus formamidopyrimidine DNA-glycosylase.* DNA Repair 5(8):947–958, 2006.

[129] Fromme, J. C., Verdine, G. L. *Structural insights into lesion recognition and repair by the bacterial 8-oxoguanine DNA glycosylase MutM.* Nature 9(7):544–552, 2002.

[130] Vriend, G. *WHAT IF: a molecular modeling and drug design program.* Journal of Molecular Graphics 8(1):52–56, 1990.

[131] Dodson, M. L., Michaels, M. L., Lloyd, R. S. *Unified catalytic mechanism for DNA glycosylases.* Journal of Biological Chemistry 269(52):32709–32712, 1994.

[132] Kow, Y. W., Wallace, S. S. *Mechanism of action of Escherichia coli endonuclease III.* Biochemistry 26(25):8200–8206, 1987.

[133] Stote, R. H., Karplus, M. *Zinc binding in proteins and solution: a simple but accurate nonbonded representation.* Proteins 23(1):12–31, 1995.

[134] Banerjee, A., Yang, W., Karplus, M., Verdine, G. L. *Structure of a repair enzyme interrogating undamaged DNA elucidates recognition of damaged DNA.* Nature 434:612–618, 2005.

[135] Klimasauskas, S., Kumar, S., Roberts, R. J., Cheng, X. *HhaI methyltransferase flips its target base out of the DNA helix.* Cell 76(2):357–369, 1994.

[136] Goedecke, K., Pignot, M., Goody, R. S., Scheidig, A. J., Weinhold, E. *Structure of the N6-adenine DNA methyltransferase M.TaqI in complex with DNA and a cofactor analog.* Nature 8(2):121–125, 2001.

[137] Slupphaug, G., Mol, C. D., Kavli, B., Arvai, A. S., Krokan, H. E., Tainer,

J. A. *A nucleotide-flipping mechanism from the structure of human uracil-DNA glycosylase bound to DNA.* Nature 384:87–92, 1996.

[138] Yamagata, Y., Kato, M., Odawara, K., Tokuno, Y., Nakashima, Y., Matsushima, N., Yasumura, K., Tomita, K., Ihara, K., Fujii, Y., Nakabeppu, Y., Sekiguchi, M., Fujii, S. *Three-dimensional structure of a DNA repair enzyme, 3-methyladenine DNA glycosylase II, from Escherichia coli.* Cell 86(2):311–319, 1996.

[139] Cheng, X. *Structure and function of DNA methyltransferases.* Annual Review of Biophysics and Biomolecular Structure 24:293–318, 1995.

[140] Mol, C. D., Parikh, S. S., Putnam, C. D., Lo, T. P., Tainer, J. A. *DNA repair mechanisms for the recognition and removal of damaged DNA bases.* Annual Review of Biophysics and Biomolecular Structure 28:101–128, 1999.

[141] Roberts, R. J., Cheng, X. *Base flipping.* Annual Review of Biochemistry 67:181–198, 1998.

[142] Schneider, T. D. *Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation.* Nucleic Acids Research 29(23):4881–4891, 2001.

[143] Englander, S. W., Kallenbach, N. R. *Hydrogen exchange and structural dynamics of proteins and nucleic acids.* Quarterly Reviews of Biophysics 16(4):521–655, 1983.

[144] Kochoyan, M., Leroy, J. L., Guron, M. *Proton exchange and base-pair lifetimes in a deoxy-duplex containing a purine-pyrimidine step and in the duplex of inverse sequence.* Journal of Molecular Biology 196(3):599–609, 1987.

[145] Kochoyan, M., Lancelot, G., Leroy, J. L. *Study of structure, base-pair opening kinetics and proton exchange mechanism of the d-(AATTGCAATT) self-complementary oligodeoxynucleotide in solution.* Nucleic Acids Research 16(15):7685–7702, 1988.

[146] Leroy, J. L., Charretier, E., Kochoyan, M., Guron, M. *Evidence from base-*

pair kinetics for two types of adenine tract structures in solution: their relation to DNA curvature. Biochemistry 27(25):8894–8898, 1988.

[147] Guéron, M., Leroy, J. L. Nucleic acids and molecular biology: Springer Verlag. 1992: 1–22.

[148] Seibert, E., Ross, J. B. A., Osman, R. *Contribution of opening and bending dynamics to specific recognition of DNA damage.* Journal of Molecular Biology 330(4):687–703, 2003.

[149] Cao, C., Jiang, Y. L., Stivers, J. T., Song, F. *Dynamic opening of DNA during the enzymatic search for a damaged base.* Nature 11(12):1230–1236, 2004.

[150] Torrie, G. M., Valleau, J. P. *Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling.* Journal of Computational Physics 23(2):187–199, 1977.

[151] Grossfield, A. The Weighted Histogram Analysis Method (WHAM). http://membrane.urmc.rochester.edu.

[152] Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., Kollman, P. A. *Multidimensional free-energy calculations using the weighted histogram analysis method.* Journal of Computational Chemistry 16(11):1339–1350, 1995.

[153] Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., Kollman, P. A. *The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method.* Journal of Computational Chemistry 13(8):1011–1021, 1992.

[154] Giudice, E., Várnai, P., Lavery, R. *Base pair opening within B-DNA: free energy pathways for GC and AT pairs from umbrella sampling simulations.* Nucleic Acids Research 31(5):1434–1443, 2003.

[155] Banavali, N. K., MacKerell, A. D. *Free energy and structural pathways of base flipping in a DNA GCGC containing sequence.* Journal of Molecular Biology 319(1):141–160, 2002.

[156] Barthel, A., Zacharias, M. *Conformational transitions in RNA single uridine and adenosine bulge structures: a molecular dynamics free energy simulation study.* Biophysical Journal 90(7):2450–2462, 2006.

[157] Ravindranathan, K. P., Gallicchio, E., Levy, R. M. *Conformational equilibria and free energy profiles for the allosteric transition of the ribose-binding protein.* Journal of Molecular Biology 353(1):196–210, 2005.

[158] Giudice, E., Várnai, P., Lavery, R. *Energetic and conformational aspects of A:T base-pair opening within the DNA double helix.* ChemPhysChem 2(11):673–677, 2001.

[159] Huang, N., Banavali, N. K., MacKerell, A. D. *Protein-facilitated base flipping in DNA by cytosine-5-methyltransferase.* Proceedings of the National Academy of Sciences of the United States of America 100(1):68–73, 2003.

[160] Ramstein, J., Lavery, R. *Energetic coupling between DNA bending and base pair opening.* Proceedings of the National Academy of Sciences of the United States of America 85(19):7231–7235, 1988.

[161] Kleywegt, G. J., Jones, T. A. *Detection, delineation, measurement and display of cavities in macromolecular structures.* Acta Crystallographica Section D Biological Crystallography 50(Pt 2):178–185, 1994.

[162] Arnold, K., Bordoli, L., Kopp, J., Schwede, T. *The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling.* Bioinformatics 22(2):195–201, 2006.

[163] Golan, G., Zharkov, D. O., Feinberg, H., Fernandes, A. S., Zaika, E. I., Kycia, J. H., Grollman, A. P., Shoham, G. *Structure of the uncomplexed DNA repair enzyme endonuclease VIII indicates significant interdomain flexibility.* Nucleic Acids Research 33(15):5006–5016, 2005.

[164] Schlitter, J., Engels, M., Krüger, P. *Targeted molecular dynamics: a new approach for searching pathways of conformational transitions.* Journal of Molecular Graphics 12(2):84–89, Jun 1994.

[165] Várnai, P., Lavery, R. *Base flipping in DNA: pathways and energetics stud-*

ied with molecular dynamic simulations. Journal of the American Chemical Society 124(25):7272–7273, 2002.

[166] Hunter, C. A., Sanders, J. K. M. *The nature of π-π interactions.* Journal of the American Chemical Society 112(14):5525–5534, 1990.

[167] Tchou, J., Michaels, M. L., Miller, J. H., Grollman, A. P. *Function of the zinc finger in Escherichia coli Fpg protein.* Journal of Biological Chemistry 268(35):26738–26744, 1993.

[168] Kropachev, K. Y., Zharkov, D. O., Grollman, A. P. *Catalytic mechanism of Escherichia coli endonuclease VIII: roles of the intercalation loop and the zinc finger.* Biochemistry 45(39):12039–12049, 2006.