Exploring DNA Damage and Repair through Molecular Dynamics Simulations

A Dissertation Presented

by

Arthur John Campbell

to

The Graduate School

In Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Chemistry

Stony Brook University

May 2011

Stony Brook University

The Graduate School

Arthur John Campbell

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Carlos Simmerling, Ph.D., Advisor
Department of Chemistry, Stony Brook University

Orlando Schärer, Ph.D., Chairperson of Defense
Departments of Pharmacological Sciences and Chemistry, Stony Brook University

Jin Wang, Ph.D., Third Member
Department of Chemistry, Stony Brook University

Miguel Garcia-Diaz, Ph.D., Outside Member
Department of Pharmacological Sciences, Stony Brook University

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Exploring DNA Damage and Repair through Molecular Dynamics Simulation

by

Arthur John Campbell

Doctor of Philosophy

In

Chemistry

Stony Brook University

2011

Everyday our cells are exposed to various forms of endogenous and exogenous DNA damaging agents. One of the most cytotoxic forms of endogenous damage comes from Reactive Oxygen Species. Oxidative DNA damage has been linked to a number of human diseases, including cancer. The most common oxidative lesion found in DNA is 8-oxo-guanine, or 8-oxo-G. It is known that formamidopyrimidine DNA glycosylase (Fpg) distinguishes between undamaged guanine bases and damaged 8-oxo-G, however, the mechanism by which this discrimination is carried out is currently under debate. There are a number of X-ray crystallographic structures available that highlight Fpg's damage recognition and base eversion process that involves the repair of these lesions, yet there are some limitations to these experiments. These experimental structures are only snapshots along the reaction coordinate, therefore more detail is needed to understand Fpg's mechanism between these snapshots. In this

work, computer modeling was used to measure the eversion process as well as pinpoint key components involved in damage recognition of 8-oxo-G.

A form of exogenous damage of particular interest in this work is the cancer chemotherapeutic agent Nitrogen Mustard (NM). NM is known to form various crosslinks in DNA. The crosslink that has been shown to be the most effective in inhibiting cell growth and facilitating apoptosis is the 1,3 interstrand crosslink formed in a 5' GpNpC sequence. Little structural information is known about this crosslink because it is difficult to study experimentally, as it is a promiscuous agent forming various types of crosslinks that are prone to depurination. The Schärer laboratory has created a protocol to synthesizing a stable analog of NM but it is unknown if this analog is a good mimic of the native NM adduct. In this work, molecular modeling was used to compare the natural NM to its chemically stable analog to validate the use of these analogs. These crosslinks were also compared to undamaged DNA to quantify the amount of distortion created by a 1,3 NM crosslink.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	**Full Form of the Abbreviation**
Å	Angstrom
A	Adenine
AMBER	Assisted Model Building with Energy Refinement
bp	Base Pair
BER	Base Excision Repair
C	Carbon
C	Cytosine
COM	Center of Mass
Fpg	Formamido pyrimidine DNA glycosylase
GB	Generalized Born
G	Guanine
ICL	Interstrand Crosslink
LJ	Lennard-Jones
MD	Molecular Dynamics
MM	Molecular Mechanics
NER	Nucleotide Excision Repair
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
PHI	Measure of Base Eversion
PME	Particle Mesh Ewald
PMF	Potential of Mean Force
RMSD	Root Mean Squared Deviation
SASA	Solvent Accessible Surface Area
TIP3P	Transferable Intermolecular Potential 3 Points
T	Thymine
VDW	van der Waals
WC	Watson Crick
WHAM	Weighted Histogram Analysis Method
8-oxo-G	8-oxo-Guanine

Acknowledgements

First I would like to thank my advisor, Professor Carlos Simmerling whose work ethic and enthusiasm for molecular modeling is unmatched. He is by far the hardest worker I have ever seen. I am extremely grateful for the ability he has given to me to solve complex scientific problems. I would also like to thank him for giving me the opportunity to work for him these past years.

I would like to thank my co-advisor Professor Orlando Schärer who has enlightened me on all different facets of DNA damage and repair. His enthusiasm for DNA repair is contagious and I truly have caught the bug.

I would sincerely like to thank Professor Arthur Grollman and Professor Carlos de los Santos both of whom have given me great guidance over the years for the DNA damage project.

I would also like to thank the Integrative Graduate Education and Research Traineeship (IGERT) program which has generously provided me with funding and has given me the opportunity to collaborate with several groups which normally I would not have had the opportunity to. In particular, I would like to thank Professor Martin Schoonen and Professor Stella Tsirka.

Sincere thanks to my third committee member Professor Jin Wang who has given me valuable insights into my work over the years. I would also like to thank my outside committee member Professor Miguel Garcia-Diaz who has an incredible enthusiasm for molecular modeling and bio-molecular structure. I would like to thank Professor Robert Rizzo for giving me great feedback on my third meeting proposal.

I would also like to thank my dear friend and experimental collaborator on the interstrand crosslink project, Dr. Angelo Guainazzi. We had a great time discussing our successes and failures throughout our years at Stony Brook.

I would also like to thank all my friends at Stony Brook University that have made the PhD. experience a great one. Specifically, I would like to say thanks to all my friends in the Simmerling and Schärer laboratories that have made science so much fun over the years. In particular I would like to thank Amber, Asim, Barbara, Banke, Carl, Christina, Ding, Lauren, Lidija, Kun, Salma, Tony and Vinh. True friends are hard to come by and I have been very fortunate to have you guys in my life.

I would also like to thank my family, especially my mother for all of her encouragement in pursuing higher education over the years. Lastly I would like to thank my beautiful wife Alison without whom, none of this would have been possible or so much fun.

Chapter 1 Introduction

1.1 DNA damage and Repair

1.1.1 DNA damage and its associated health concerns

Every day, the human genome is exposed to various forms of exogenous and endogenous DNA damaging agents. It is vital for human survival that our genomic information remains intact, however there are an enormous number of known toxic agents that cause DNA damage.[1] These DNA-damaging agents can have a cause as trivial as eating burnt meat (benzopyrene)[2], or might arise from a more complex origin, such as the use of Chinese herbal medicines for weight loss (aristolochic acid)[3]. Some of these DNA damaging agents can disrupt normal cellular activities. This disruption of normal cellular activities may be due to the slowing of replication or transcription until the lesions can be corrected. Certain lesions can even form irreversible mutations in the genome that can lead to abnormal cellular behavior or even cell death. Irreversible mutations can lead to various diseases for the organism, such as cancer.[4-6] Because the human genome encodes for all cellular components except the mitochondria, compromising the genome has far-reaching consequences. Certain components of the genome, such as proteins, may not fold properly and interactions with other proteins may be disrupted, thus causing cellular pathways to be compromised. One example of this is the p53 protein, which is involved in various forms of regulation in the cell such as cell growth and apoptosis.[6,7] It has been shown that ~50% of human cancers have mutations in the p53 gene.[6,7] Thus, it is vital that the genomic integrity stays intact for proper cellular and organism function.

The end products of two DNA damaging agents are of particular interest in this work. One of these DNA-damaging agents is an exogenous DNA-damaging cancer chemotherapeutic agent known as nitrogen mustard (NM), and the other is endogenous DNA-damaging reactive oxygen species (ROS). Nitrogen mustard forms various types of crosslinks in the cell. Of particular interest is the crosslink that is the most lethal, the 1,3 interstrand crosslink (ICL).[8,9] When several 1,3 ICLs are formed from NM in a cell they can become cytotoxic as they have the ability to inhibit replication and transcription by blocking strand separation and polymerase activity.[10,11] Another form of DNA damage of interest in this work is one of the most cytotoxic forms of endogenous damage which comes from exposure to ROS. The "oxygen paradox" eloquently describes the fine line between a cell's need for oxygen and the consequence that overexposure to ROS may lead to toxicity.[1] Oxidative DNA damage has been linked to a number of human diseases including cancer.[4,5] One of the most common oxidative lesions formed in DNA is 8-oxo-guanine (8-oxo-G) which is of particular interest in this work.[12-14] In this chapter we will discuss the basic properties of DNA and how exposure to these two toxic elements, reactive oxygen species and nitrogen mustard, can disturb these properties.

1.1.2 Properties of DNA

Each nucleotide in deoxyribonucleic acid (DNA) is composed of three basic elements: a phosphate (PO_4^{-2}), sugar (deoxyribose), and nucleobase group (Figure 1-1, highlighted red, yellow and blue respectively).[15] Normal DNA comprises four nucleobases- adenine (A), guanine (G), cytosine (C), and thymine (T) (Figure 1-2). These four nucleobases are further divided into two categories based on their cyclic rings: adenine and guanines are classified as purines, and cytosine and thymine are classified as pyrimidines (Figure 1-2, highlighted blue and yellow respectively). Each base has its own Watson-Crick (WC) hydrogen bonding partner (also known as a base pair); guanine hydrogen bonds to a cytosine while adenine hydrogen bonds to a thymine (Figure 1-3). The different WC faces of each nucleobase make it sterically and electrostatically impossible for a nucleobase to form a WC hydrogen bond with any other nucleobase other than its bonding partner, thus increasing the fidelity of replication.[16] In a G:C base pair, three WC hydrogen bonds are formed, while in an A:T base pair two WC hydrogen bonds are formed. When these base pairs are formed in concert under normal conditions they create a double helix (Figure 1-4).

A DNA double helix consists of two antiparallel strands of DNA that form a cork-screw-like spiral. The genetic information (each base pair) is located in the central region of the helix and is surrounded by the sugars as well as the phosphate backbone. The direction of the anti-parallel strands is often labeled as either going in the 5' to 3' or 3' to 5' direction. This labeling refers to the orientation of the 3' and 5' deoxyribose sugar groups on each nucleobase. Under normal temperature, pressure, and ionic concentration conditions, double-stranded DNA forms B-form DNA (Figure 1-4). In B-form DNA, there are two distinct grooves that form: the minor and major grooves. Besides having overall differences in the widths and depths of the two grooves the chemical structures available for contacts from each nucleobase are different in each groove (Figure 1-3). It has previously been shown through X-ray crystallography as well as molecular modeling that the sequence of DNA has a direct impact on the overall structure.[17,18] For example, duplex DNA with A:T tracks has been shown to be wound tighter then sequences that contain mixtures of G:C and A:T tracks.[17,18] It has been shown that a distinct spine of hydration forms in the minor groove of DNA, particularly with A:T tracks.[19] Sequence dependence of these ordered waters was also confirmed with molecular dynamics simulations.[20] The spine of hydration may aid in certain protein-DNA interactions.[20]

The effects of DNA damage can be understood at the molecular level by comparing the normal properties that are found in DNA to DNA that has been damaged. Under normal conditions DNA is not a perfectly conical duplex as shown in Figure 1-4, but is wrapped neatly around histones and tightly packaged into nucleosomes, in order to save precious space in the nucleus.[21] While it would be more accurate to study the structural properties of DNA when incorporated into the nucleosome, it is currently difficult to accomplish this experimentally or computationally. Global genome repair (GGR) is known to repair damaged DNA in the nucleosome but experimental laboratories do not currently study this type of DNA repair due to the difficulty of isolating a nucleosome in vitro. In studying DNA damage using molecular modeling, approximately 11 base pairs are usually selected to model duplex DNA. The rationale behind selecting approximately 11 basepairs is that 10.5 base pairs make a complete turn in DNA and the pi-pi base stacking interactions further stabilize the duplex.[22] This trend is also observed in X-ray crystallography, where it is more common to crystallize a duplex with 11 or 12 base pairs than with fewer of them.[23-26]

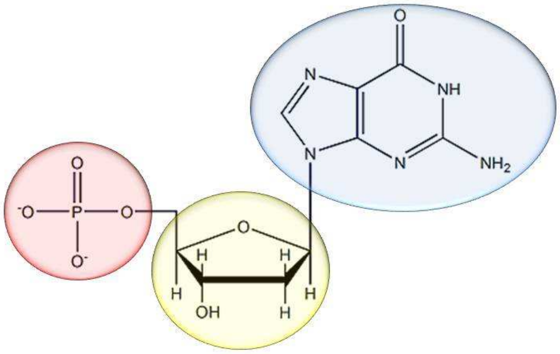


Figure 1-1. An undamaged nucleotide comprises a phosphate (PO_4^{-2}, red), deoxyribose sugar (yellow), and a nucleobase (blue). The bond connecting the sugar to the base is referred to as the glycosidic bond.

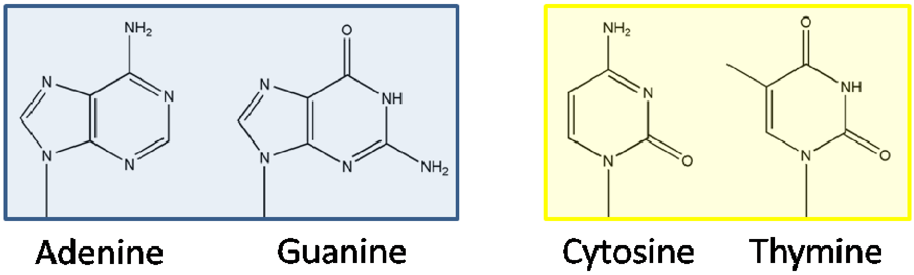


Figure 1-2. The four standard deoxyribonucleic acid nucleobases. The two bases highlighted in blue are the purine bases adenine and guanine. The two bases highlighted in yellow are the pyrimidine bases cytosine and thymine.

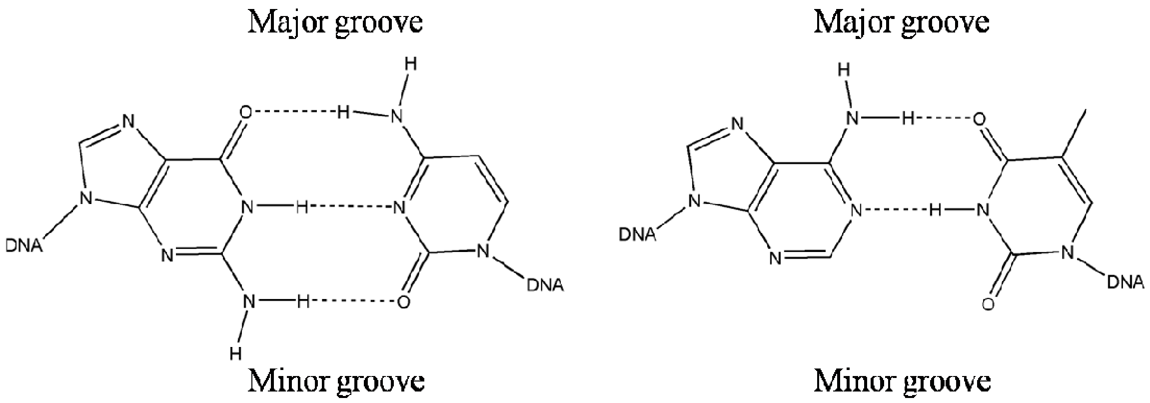


Figure 1-3. Watson-Crick base pairs G:C (left) and A:T (right). Watson-Crick hydrogen bonds are shown as dashed lines.

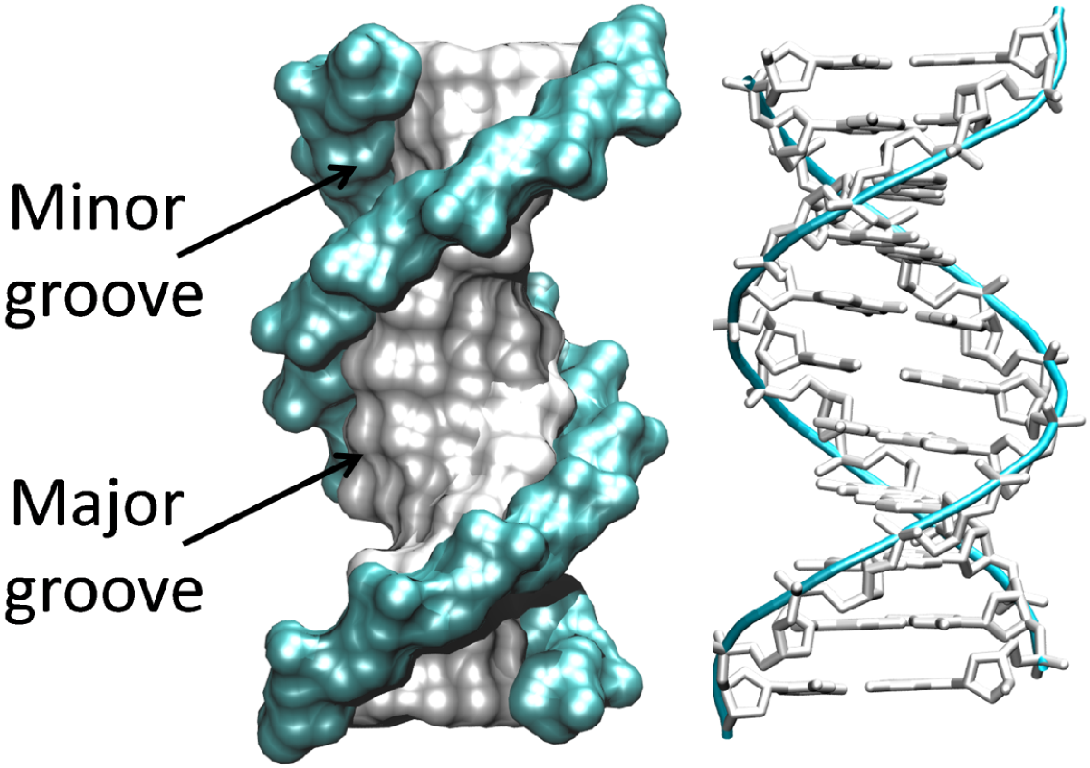


Figure 1-4. Structure of B-form duplex DNA. The backbone is highlighted in cyan. The left structure is in msms representation and the right structure is in licorice. These structures were created in Nucgen.[27] The major and minor grooves are labeled.

1.1.3 DNA repair in BER and NER

Elaborate cellular repair processes have evolved to combat various types of DNA damage which encompasses base excision repair (BER) and nucleotide excision repair (NER). In NER several proteins work in a concerted approach to repair an array of helix distorting lesions in DNA. This process recognizes large-scale distortions in DNA, and does not target individual damaged nucleobases. Additionally this process places a large energetic demand on the cell compared to repair pathways such as BER, as NER depends on the helicases XPB and XPD to unwind the DNA, a process which consumes ATP.[28,29] In BER, individual glycosylases excise specific lesions that are generally more common than the lesions targeted in NER.[1] The repair processes of BER are done with high fidelity and little or no energetic consumption.[1]

Although BER and NER are different repair pathways with different mechanisms to correct the damaged site, they share a common problem- the site of interest must first be recognized and distinguished from nondamaged DNA. In NER, XPC/Rad23B recognizes the damaged region and initiates a cascade of NER-dependent events such as damage verification and helix unwinding followed by damage removal and strand replacement.[30] In BER, the pathway is less complex; individual glycosylases recognize and remove specific types of damage. The processes of BER are executed with little or no energetic consumption.

A variety of theories exist on how damage is initially recognized by repair proteins. It has been hypothesized by Stivers et al that uracil DNA glycosylase uses a combination of frequent hopping and short distances of sliding to find an extrahelical uracil.[31] The Verdine group has speculated that the BER glycosylase, hOGG1, uses fast sliding coupled with less frequent hopping to find intrahelical damage along sections of DNA searching for 8-oxo-G.[32-34] It is difficult to draw concrete conclusions from these studies since these laboratories do not account for DNA being wrapped around histones and tightly packaged in the nucleosome which would sterically block these repair proteins from completely encircling the DNA while sliding. Different repair proteins may also vary their scanning and hopping depending on the type of damage and whether the damaged nucleobase(s) are extrahelical or intrahelical as in the two examples discussed here for Uracil DNA glycosylase and hOGG1 respectively.

A large amount of X-ray crystallographic data exists revealing a variety of protein-DNA interactions. Proteins involved in DNA binding often share common motifs such as the helix-hairpin-helix (HhH) motif.[35] Most BER glycosylases share the HhH motif such as MagIII [36], MutY[37], MIG[38], AlkA[39], OGG[40] and TagA[41]. DNA-binding proteins of other repair pathways also share the HhH motif such as the 5' endonuclease ERCC1/XPF which plays a role in 5' incision in NER.[42]

Besides sharing similar motifs, BER glycosylases share common mechanisms to remove damaged DNA. Upon initial binding, the glycosylase may evert the nucleobase of interest out of the duplex for excision. It is currently unknown how these enzymes differentiate damage from nondamaged DNA, which is one of the topics of interest in this work. As previously mentioned, these glycosylases share similarities in their folds as well as certain functions, and it is not unreasonable to suggest that a mechanism for discrimination may have similarities amongst different glycosylases.

Numerous computational studies have used molecular dynamics to investigate protein-DNA complexes in which a base is everted out of duplex DNA as well as in naked DNA.[43-52] Several definitions have been proposed to help to quantify base eversion.[43,45,53-55] One of these definitions has been used to study base eversion of cytosine in duplex DNA when complexed to a methyltransferase glycosylase. Broyde et al have studied the energy required to evert a nucleobase containing benzopyrene from a single strand of DNA to better understand how the nucleobase can be recognized by XPC/rad23b.[52]

1.1.4 Exogenous DNA damage caused from nitrogen mustard

Cells are under constant bombardment by naturally- and unnaturally-occurring foreign agents that potentially lead to cell death. Many of these agents attack the nucleic acid of the cell, where they may block replication and transcription events and/or cause mutagenicity. Eukaryotic cells have evolved many mechanisms that work on the excision and repair of these malignant lesions. One particularly interesting cytotoxic agent is nitrogen mustard (NM) (Figure 1-5), commonly known as mustard gas, which forms interstrand crosslinks that are covalent bridges between two complementary strands of DNA. These crosslinks inhibit essential processes such as DNA replication and transcription. Nitrogen Mustards and their analogs are a cytotoxic class of bifunctional alkylating agents that form various DNA crosslinks.[8,9] Among the adducts that form from the reaction of these compounds with DNA, DNA interstrand crosslinks (ICLs) cause the greatest cytotoxicity to the cell. ICLs form covalent bridges between two complementary strands of DNA, thereby inhibiting essential processes such as DNA replication and transcription.

Nitrogen mustards are known to form ICLs between two guanines of complementary strands, (5'GpNpC) of DNA (Figure 1-6). It has previously been shown that the instability of the glycosidic bond on the guanines makes the ICL of NM particularly unstable for experimental investigation.[8,56] In particular, this instability is due to the increased likelihood of depurination occurring to the crosslinked bases.[57] The increase likelihood of depurination is a particular problem when trying to experimentally test the ICL since the crosslink of interest will be altered. Although it is difficult to study the 1,3 ICL experimentally, depurination occurs on a slower timescale than that of the quickly replicating cancerous cells; therefore it is an ideal agent for treatment of cancer cells but not an ideal candidate for experimental studies.

Because NM is a promiscuous crosslinking agent, experimental treatment of cells leads to numerous types of crosslinks at different locations: interstrand crosslinks, intrastrand crosslinks, and DNA-protein crosslinks. It is therefore difficult to study defined crosslinks caused by NM. To address this issue, the Schärer group has developed a synthetic approach for the generation of NM analog ICLs by the incorporation of aldehyde precursors on two complementary DNA strands and the use of a specific double reductive amination reaction to yield the crosslink.[57] In these structural analogs the N7 positions of both guanines have been replaced with carbons, giving enough stability to prevent depurination and allow experimental investigation.[57] This methodology provides an opportunity for the generation of ICLs which vary in the length of the linker connecting the two DNA strands by using a variety of amines in the reductive amination reaction. This process can be used to study how various linker lengths are repaired in the cell.

Currently, there are no crystallographic structures of nitrogen mustard in the biologically relevant 5' GpNpC position, however there are structural analogs of ICLs in DNA in different positions, such as the 1,2 GpC crosslink in which two cytosines are crosslinked across the major grooves.[58] The 1,2 GpC crosslink does show minor local damage around the ICL. Molecular dynamics simulations have previously been used to simulate nitrogen mustard in the 1,3 GpNpC position where the crosslink was built into B-form DNA.[59,60] The basis of this computational work was to compare 1,2 and 1,3 ICLs that were formed from nitrogen mustard in duplex DNA. These simulations found that the crosslink brought the two strands of DNA closer than DNA with no crosslink.[59,60] These simulations had several limitations: the computational power available when the study was undertaken 15 years ago only allowed the investigators to simulate the DNA for 200ps, and these crosslinks were built in as a single residue, thus creating abnormally long bond lengths which were needed to relax the structure but may have caused some artificial deformation.

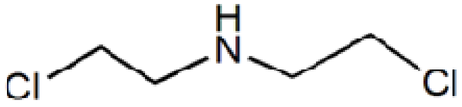


Figure 1-5. Chemical structure of nitrogen mustard (NM)

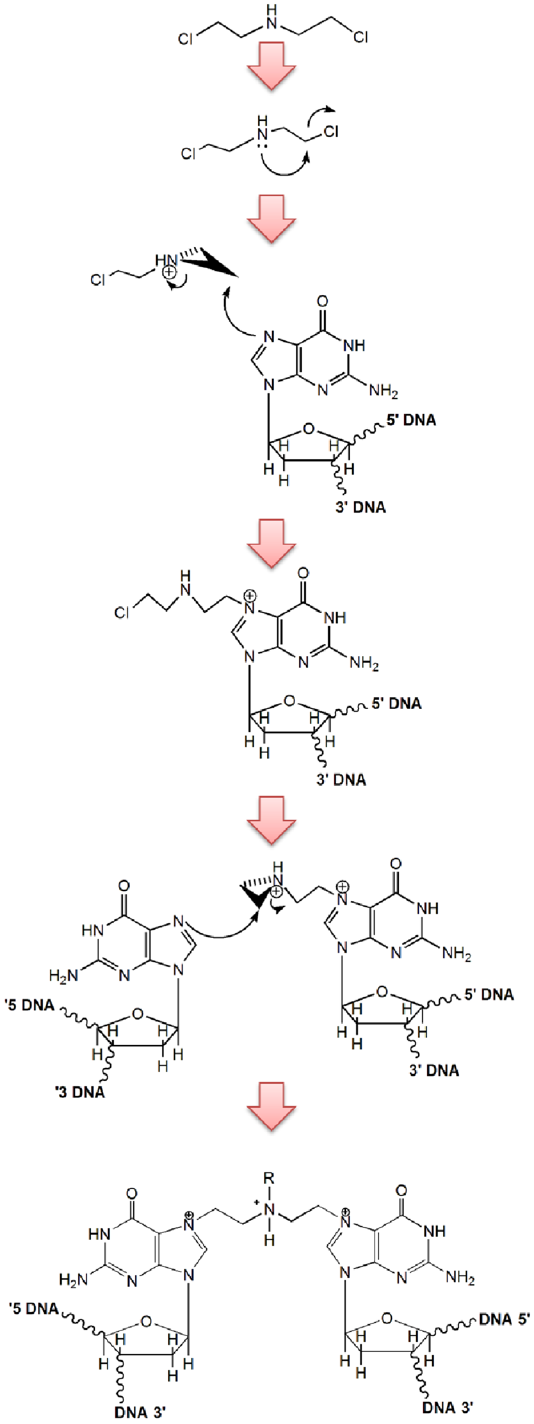


Figure 1-6. Mechanism for the chemical formation of a G to G interstrand crosslink by nitrogen mustard.[61]

1.1.5 Endogenous DNA damage caused from ROS on DNA

A fine line exists between a cell’s need for oxygen and the consequence that overexposure to reactive oxygen species may lead to toxicity.[1] A number of human diseases, including cancer, have been linked to oxidative DNA damage. The most common oxidative lesion found in DNA is 8-oxoguanine (8-oxo-G). Correcting the 8-oxo-G lesion in DNA is essential to blocking the associated diseases that are a consequence of not repairing it; therefore, a great deal of work has been done studying this lesion and how it is corrected.[62-66] A better understanding of repair proteins that remove 8-oxo-G from DNA such as the bacterial glycosylase formamidopyrimidine DNA glycosylase (Fpg), which is of particular interest in this work, may lead to a more thorough understanding of how glycosylases work in general to repair damaged DNA and prevent associated diseases. Finding and excising an 8-oxo-G lesion in a sea of undamaged bases may appear to be a daunting task, but the Fpg glycosylase can accomplish this task efficiently and without the need for energetic consumption.

There is a minor chemical difference of two atoms between an undamaged G and 8-oxo-G (Figure 1-7). Although this difference is small, it reverses the electrostatics on the major groove face of the 8-oxo-G nucleobase compared to that of G, allowing it to preferentially favor hydrogen bond formation to an adenine upon a single round of replication in a Hoogsteen base pairing manner as opposed to a Watson-Crick manner (Figure 1-8).

The catalytic mechanism of how Fpg excises 8-oxo-G has been well characterized (Figure 1-9).[67,68] Fpg carries out a nucleophilic attack on the C1' of the deoxyribose sugar of 8-oxo-G by proline 1 when in the active site, followed by β and δ elimination.[67-69] In the active site, 8-oxo-G is held in place by several hydrogen bonds to the O6 of the nucleobase.[70] A crystallographic structure of 8-oxo-G positioned in the active site loop was first demonstrated by the Verdine group (PDB 1R2Y).[70] The 1R2Y structure was trapped while probing 8-oxo-G in the extrahelical position by making an E2Q mutation, making the enzyme catalytically inactive.

Fpg has been shown through X-ray crystallography to probe for damage using an aromatic wedge (Phe 113) as well as two residues to plug the gap (Arg 112 and Met 76) and recognize the orphan cytosine opposite 8-oxo-G (PDB 2F5O, 3GQ3).[71,72] Because neither Phe 113, Arg 112 nor Met76 comes into direct contact with the damaged side of 8-oxo-G when complexed to the DNA in the intrahelical position, there must be an indirect readout mechanism that can recognize damage. An indirect readout mechanism suggests that the enzyme can sense the damaged nucleobase without directly contacting the damage. In the 2F5O structure G, is in the intrahelical position and is held in place by a crosslink to Fpg at residue 166.[71] The 3GQ3 structure is crosslinked at residue 166 as well, but also has the active loop deleted allowing 8-oxo-G to stay in the intrahelical position.[72] The deletion of the loop suggests that the active site loop may play a role in helping 8-oxo-G evert out of the duplex. To test this hypothesis, Verdine also had crystallized several structures of 8-oxo-G in duplex DNA complexed to Fpg, using a crosslink at residue 166 where point mutations were made in the active site loop, confirming that some residues in the loop play a role in facilitating eversion (PDB's 3GP1, 3GPP).[72] In another paper by Karplus and Verdine, a structure of Fpg-DNA was captured by crosslinking N174C.[73] This structure contains a G flipped out of the duplex and in the minor groove suggesting that base eversion occurs through the minor groove (PDB's 3JR4).[73]

Karplus and Verdine used several computational techniques to measure if eversion is more favorable for 8-oxo-G over G through the minor groove.[72] They found that it is indeed more favorable to evert 8-oxo-G (~13kcal/mol) compared to G (~18kcal/mol).[72] Although this work does suggest that it is more favorable for 8-oxo-G to evert than G, it does not address how Fpg, which consumes no energy can overcome an energetic barrier of ~13kcal/mol.[72]

A series of pre-steady-state kinetic studies using time course Trp fluorescence experiments on the *Eco*-Fpg-F113W system were previously studied.[74-76] These stop-flow fluorescence experiments were used to better characterize whether an indirect mechanism exists that can help distinguish between damaged (8-oxo-G) and undamaged (G). These fluorescence experiments show that Fpg facilitates eversion of both 8-oxo-G and G out of duplex DNA to transiently stable states. These experiments do not give any directionality as to which way these two nucleobases evert nor do they suggest what is playing a role in stabilizing them.

The indirect readout mechanism by which Fpg can distinguish between an undamaged guanine nucleobase and a damaged 8-oxo-G is currently unknown. As previously stated, there exists a large amount of structural data that highlight Fpg's damage recognition and base eversion processes.[66,70-72] These experimental structures are only snapshots along the process, but we would like to understand what happens between these snapshots, more detail is needed to elucidate the steps of this process. In this work, computer modeling was used to measure the eversion process of Fpg and pinpoint key components involved in damage recognition of 8-oxo-G.

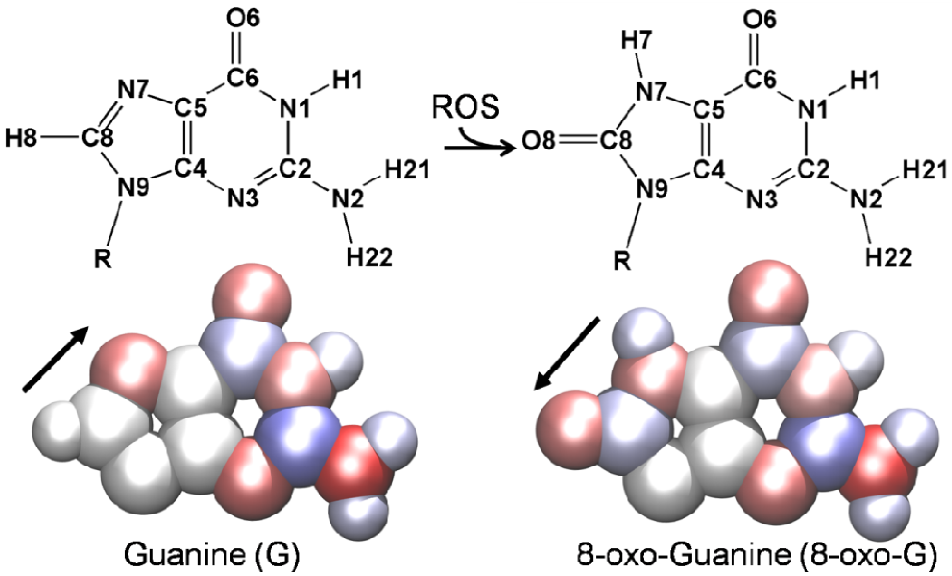


Figure 1-7 Atom names of guanine (top left) and 8-oxo-guanine (top right). The chemical differences between the two nucleobases can be compared by their electrostatic differences: bottom left (G) and bottom right (8-oxo-G). The colors used for the electrostatics representation are red, white, and blue for electronegative, neutral, and electropositive, respectively. Deeper variations of red and blue signify a greater extent of electronegativity or electropositivity respectively. Arrows indicate the change of the local dipole moment around the damage site (pointing towards the more negative direction).

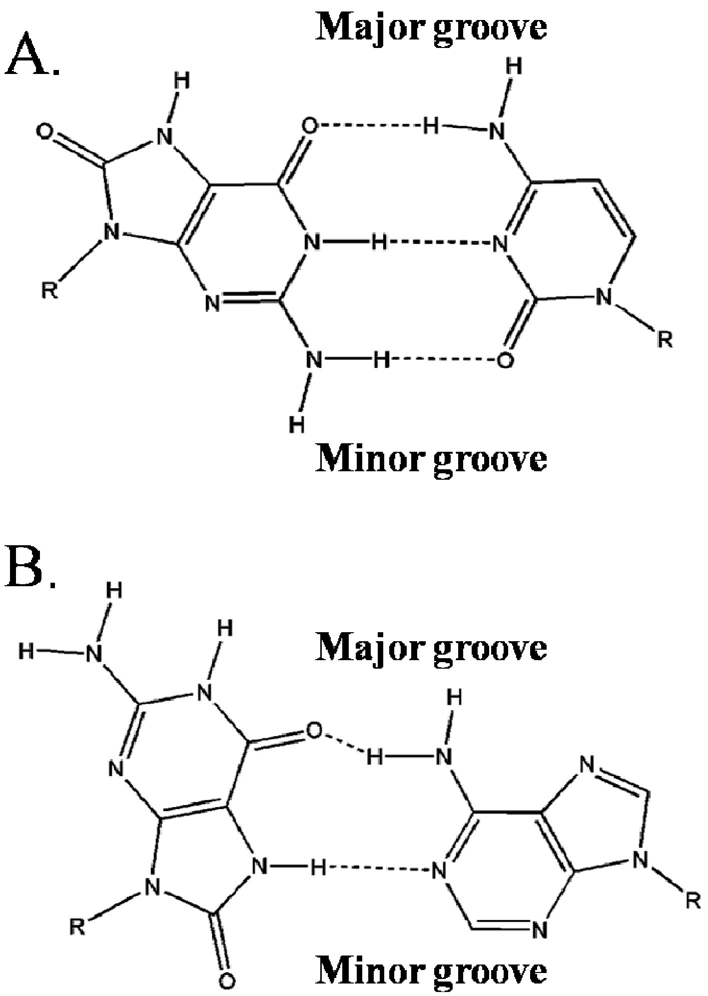


Figure 1-8 A.) Watson-Crick hydrogen bond formation between an 8-oxo-guanine and cytosine B.) Hoogsteen base pair formation between an 8-oxo-guanine and an adenine.

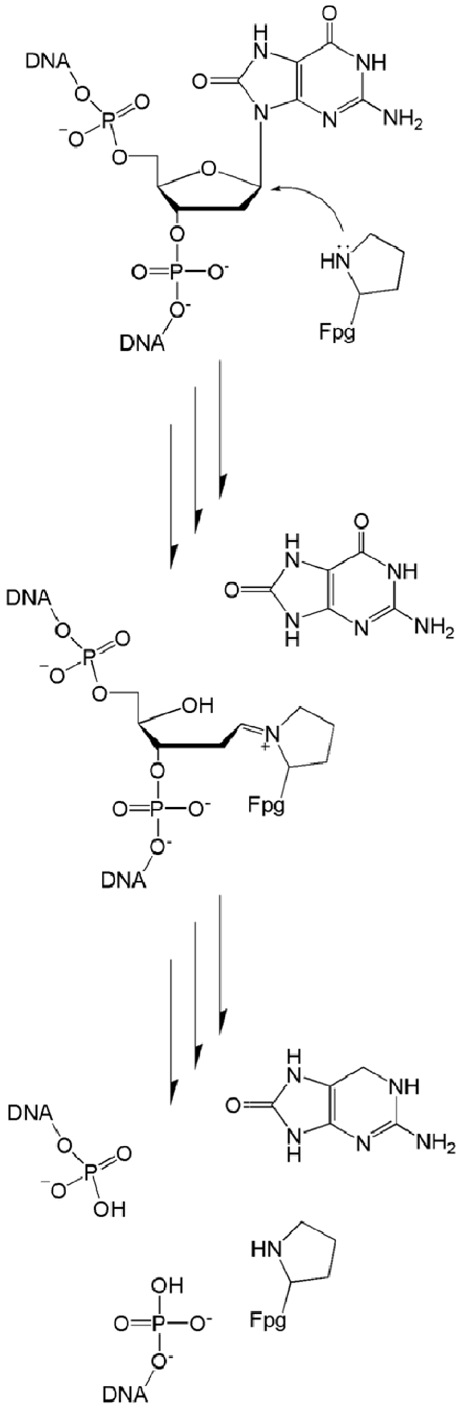


Figure 1-9 The proposed catalytic mechanism for the removal of 8-oxo-G by Fpg when in the active site loop of DNA. In the active site, proline 1 carries out a C1' attack on deoxyribose of 8-oxo-G (top), creating a Schiff base intermediate (middle). From the intermediate, β (beta) and δ (delta) elimination occurs creating a nick in the DNA (bottom).[67-69]

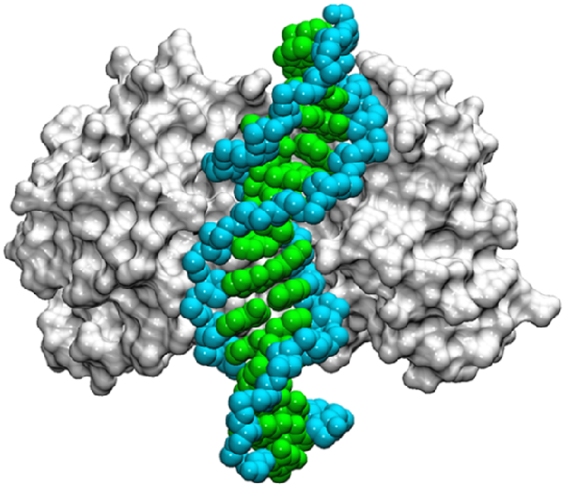


Figure 1-10 Equilibrated structure of Fpg complexed to duplex DNA (PDB 2F5O) where G was computationally mutated to an 8-oxo-G in the intrahelical position.[71]

1.2 Simulation methodology

Molecular dynamics (MD) simulations can be used to computationally study a wide range of scientific problems. Structural biology is a field in which MD simulations can have a significant impact. In structural biology, experimentalists are limited by the spatial resolution that can be obtained through experimental techniques, such as in X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR), and atomic force microscopy (AFM). Although highly successful at obtaining structural details, these experiments are limited by the fact that the structures obtained are single snapshots of highly dynamic bio-molecules, while the less populated or unstable states between those snapshots are not realized. MD simulations can be used to understand these less-populated structures between the static experimental snapshots. By better understanding the dynamic properties of these bio-molecules, structural events or processes may be elucidated. Molecular dynamics is a particularly powerful tool, as molecular interactions may be visualized from the trajectories created from the Amber simulation package as time progresses.[27] These time-dependent trajectories can be viewed with visualization software such as Visual Molecular Dynamics (VMD) or PyMol.[77] An additional benefit to using the Amber software is that it is distributed with an analysis package, Ambertools 4.1.[27] This analysis software allows the user to perform a detailed analysis of the trajectories created at atomic resolution.

1.2.1 Force fields

The fundamental component of classical molecular mechanics calculations is the force field. The force field is a potential energy function that represents the physical interactions of the molecule(s) being simulated. By calculating these physical interactions at each timestep, the potential energy of the system can be calculated and the dynamics examined as time evolves. The Amber force field, which is of particular interest in this work, has terms divided into four categories: bond lengths, bond angles, dihedral angles and non-bonded interactions. The parameters for each of the interactions found in the force field are predefined. The general Amber force field is outlined in equation 1-1 below.

$$U = \sum_{bonds} \frac{1}{2} K_b (b - b_0)^2 + \sum_{angles} \frac{1}{2} K_\theta (\theta - \theta_0)^2 + \sum_{torsions} K_\phi [1 - \cos(n\phi)]$$

$$+ \sum_{i<j,\, nonbonded} \varepsilon_{ij} \left[\left(\frac{R_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{i<j,\, nonbonded} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}} \quad \text{1-1}$$

The general Amber force field equation is divided into two parts: one part encompasses the "hard degrees of freedom" such as bond lengths, angles, and torsions, while the other part describes the nonbonded interactions, namely, the electrostatic interactions and the van der Waals forces. The summation of all five of these interactions gives the potential energy (U) of the system. The first term in the summation defines the distance between two covalently bound atoms. The second term defines the angle between three sequentially covalently bonded atoms. These two terms are described by a harmonic potential, with K_b and K_θ being their respective force constants, and (b-b_0) and (θ-θ_0) being the deviation from their respective equilibrium values. The third term defines the rotation about a bond as described by the dihedral angle. As mentioned earlier, the nonbonded component of the potential energy is divided into two parts that define the interactions between two atoms: a Coulomb potential that defines the electrostatic interactions and a Lennard-Jones potential that defines the van der Waals interactions.

The Amber all-atom force field is used to describe molecules at the atomic level. This force field is parameterized by quantum mechanical calculations as well as by fitting to experimental findings such as vibrational frequencies and structures.[78-80] The main force field used in this work is known as ff99SB with the parmbsc0 backbone additions.[78,81] The ff99SB forcefield has previously been shown to be highly accurate at modeling bio-molecules.[78] The parmbsc0 backbone parameters modify the alpha and gamma torsional terms for DNA. The parmbsc0 modified alpha and gamma torsional terms have previously been shown to give more reasonable results for simulating DNA.[81] It has been shown by comparing experimental and calculated order parameters that ff99SB is highly accurate.[82]

1.2.2 Solvation effects

Modeling bio-molecules in solvent is of interest in this work because a solvent environment represents reality more accurately than *in vacuo* simulations. It has previously been shown that solvent can have specific interactions with bio-molecules, as well as having bulk solvent effects on the system of interest.[83-85] The choice of how to model the solute-solvent interactions in molecular dynamics simulations therefore can have a direct impact on the accuracy of the system as well as the efficiency of the calculations. The two most common water models available are explicit and implicit solvent models.

In the explicit solvent water model, each water molecule is independently modeled in the system of interest. One advantage of explicitly modeling each water molecule is that specific solute-solvent interactions, such as bridging waters between an individual water molecule and two regions on the molecule of interest, can be sampled. By modeling each water molecule, the total number of atoms increases significantly and therefore the computational expense of these calculations also increases. One of the more common explicit solvent models is TIP3P, which is used throughout this work.[86] This water model was parameterized by fitting the properties of bulk water as well as its thermodynamic properties.[86] One way to reduce the computational cost of modeling in solvent is to use particle mesh Ewald (PME) to represent the long-range electrostatic interactions.[87] The coupling of explicit solvent with PME for long-range electrostatic interactions is often considered the most accurate way to explicitly model water using molecular dynamics.[88]

In the implicit solvent water model, also known as continuum solvent, the solvent is considered as a continuous medium, as opposed to the individual water molecules used in explicit solvent. There are various types of implicit solvent models available. Two largely popular models are Poisson Boltzmann (PB) and Generalized Born (GB).[89] In implicit solvent models, the free energy of solvation is estimated solely on the positions of the solute atoms.[89] By neglecting the solvent atoms in the energy calculations of the system, large computational costs can be saved. Implicit solvent models also carry certain weaknesses, as the exclusion of solvent molecules does not permit sampling of important interactions, such as bridging or ordered waters.

Explicit and implicit solvent models carry with them certain strengths and weaknesses. When modeling systems using molecular dynamics simulations, it is important to understand the overall problem that is trying to be solved and use a solvent model that will help lead to the most accurate answer. For example, various studies on modeling base eversion have used the TIP3P explicit solvent water model, as specific water-protein or water-DNA interactions may have an effect on the system as the water exposed base everts.[47,48,52,90]

1.2.3 Advanced molecular dynamics simulation methods

One of the main limitations to molecular dynamics simulations is the ability to sample biologically relevant states. To investigate biologically relevant states, enhanced sampling may be needed in order to overcome energy barriers on the molecule's potential landscape. It is therefore often beneficial to force a transition to occur during a simulation and subsequently analyze the interactions that occur as a result of this transition. There are numerous ways to force a transition to occur, such as using targeted molecular dynamics, the partial nudged elastic band method and umbrella sampling.[27,91-93]

Targeted molecular dynamics

A common way to create a transition between two states is to use a method called targeted molecular dynamics. This method uses a biasing harmonic potential to force an initial state to adopt the conformation of a targeted state. This calculation is added to the energy function of the system. Equation 1-2 gives the equation for the calculation of the targeting force in targeted molecular dynamics. K represents the force constant. N represents the number of atoms that this force will be applied to. RMSD is the difference between the current state and the targeted state. $RMSD_0$ is the targeted value.

$$E_{TMD} = K * N * (RMSD - RMSD_0)^2 \qquad 1\text{-}2$$

In targeted molecular dynamics, the initial state can be forced to move towards or away from the targeted value based on the sign of the force constant. Although this method is a useful approach to solving simple problems, it may not be able to sample complex problems, such as applying targeted molecular dynamics on the initial state. With this approach the targeted value forces the system to take the most direct spatial path regardless of whether this path is higher in energy than an indirect path. Therefore, targeted molecular dynamics may not be the best enhanced sampling technique if there is a need to sample an indirect path.

Partial nudged elastic band theory

The partial nudged elastic band (PNEB) method creates a low energy path between two or more states.[91] PNEB can be used to uncover direct or indirect paths between states, which is an advantage over methods such as targeted molecular dynamics, which cannot locate an indirect path. The PNEB code in Amber's multisander functionality allows the user to apply nudged elastic band forces to a subset of the system's atoms.[27] This method is of particular interest for large, explicitly-solvated systems, since the forces can be applied directly to the region(s) of interest. An initial path is needed before a low energy pathway can be found. To create an initial path, multiple copies of the end point structures (known as “beads”) are usually made. Alternatively, to "seed" a pathway of interest that may be indirect, midpoint bead(s) can be used to help guide the path in the direction of interest. The initial path is optimized, using several rounds of simulated annealing in which heating and cooling occur.

In PNEB, springs are used to keep each bead evenly spaced along the pathway, and a decoupling tangent based on the neighboring bead positions or energies is used in order to remove the effect of the springs during the optimization process by the forcefield.

Umbrella Sampling and WHAM analysis

As mentioned earlier, a major limitation to molecular dynamics simulations is the sampling time, which is dependent on the available computational resources. It is therefore difficult to calculate the free energy of a system between two states if the simulations cannot sample one of the states or the transition that is needed for it to occur. Ideally, if a simulated system is sampled until equilibrium is reached between the two states of interest, the free energy difference can be measured based on the populations of the two states:.

$$F(q_B) - F(q_A) = -k_B T \ln \frac{P(q_B)}{P(q_A)} \qquad 1\text{-}3$$

In this equation the free energy is calculated between population $P(q_A)$ and a reference population $P(q_B)$ at reaction coordinates A and B respectively. k_B is the Boltzmann constant and T is the temperature of the system.

When simulating a system of interest, if equilibrium between two states cannot be reached, then a method known as umbrella sampling can be used in which a biasing potential is added in order to force sampling of the region of interest.[94-96] In umbrella sampling, a group of simulations is performed while varying the biasing potential to sample between the states of interest. By applying a biasing potential, states that normally would not be sampled in unrestrained molecular dynamics simulations can be sampled. After sampling the region of interest, a potential of mean force (PMF) can be calculated by using the weighted histogram analysis method (WHAM).[94] WHAM analysis is used to unbias the data to obtain free energy profiles or PMFs.[94]

The following equation is used to measure the unbiased potential of mean force between two states when a biasing potential ($V^{(i)}$) is applied based on the reaction coordinates (*i*). K(i) is the free energy constant that is determined by equation 1-4 below:

$$F^{(i)}(q_B) = F(q_A) - kT \ln \frac{P(q_B)}{P(q_A)} - V^{(i)}(q_B) + K^{(i)} \tag{1-4}$$

$$e^{-\frac{K^{(i)}}{kT}} = < e^{-\frac{V^{(i)}(q_B)}{kT}} > \tag{1-5}$$

The underestimated constant, $K^{(i)}$ can be determined in the equation above.

$$P(q) = \sum_{i=1}^{N_w} n_i P(q)_{(i)} \times \left[\sum_{j=1}^{N_W} n_j e^{-\frac{\left[V^{(i)}(q) - K^{(i)}\right]}{kT}} \right]^{-1}$$

1-6

Equation 1-6 is the fundamental equation for WHAM analysis.[94,95] The optimized estimate of the unbiased distribution function is dependent on the reaction coordinate q weighed over the sum of the number of biased windows (N_w). The number of individual data points is given by n_i. The underestimated constant is calculated in equation 1-5.

When using the umbrella sampling method there are some variables that should be carefully considered. The fundamental assumption in umbrella sampling is that there are one or two reaction coordinates that can define the system of interest. The reaction coordinates have limitless possibilities, such as backbone dihedrals or even base eversion.[91,97] Umbrella sampling is usually used with systems that exhibit a lack of sampling in a region of the energy landscape, suggesting that there is a high barrier to enter that region, or that the region itself is significantly higher in energy. This suggests that windows would need to be restrained to stay in the region.

Another important issue is the choice of initial structures used in umbrella sampling. To create each initial window the restraint can force the reaction coordinate(s) to adopt their targeted position(s). This approach biases the system towards the initial structures. A better approach would be to couple partial nudged elastic band method with umbrella sampling.[91] First, a low energy pathway can be created between two states using partial nudged elastic band method; then structures can be extracted based on the chosen reaction coordinate(s) and used for umbrella sampling.[91]

1.2.4 Challenges and relevance to projects

In this work we are interested in modeling DNA with or without damage as well as with or without being complexed to repair enzymes. One of our interests lies in measuring the energy needed to evert a nucleobase out of DNA when bound by an enzyme. To accomplish everting a nucleobase, the partial nudged elastic band method and umbrella sampling were coupled, as this event could not be sampled using conventional unrestrained molecular dynamics. Two reaction coordinates were chosen to quantify the base eversion event, one of which was developed in our laboratory.[97] Several other laboratories have used other approaches to study base eversion such as solely using umbrella sampling or using targeted RMSD which may not sample the lowest energy path. [48,72,73]

1.3 Overview of My Research Projects

1.3.1 Modeling of interstrand crosslinks in DNA

Our interests lie in studying nitrogen mustard for several reasons. Various DNA repair pathways, including nucleotide excision repair (NER), homologous recombination, and translesion synthesis work together to repair ICLs, but the details of how repair is achieved are not understood.[8] In particular, the relationship between the structure of the ICL and its repair are not known. The ability to predict and characterize how ICLs influence DNA structure will provide opportunities to address this issue. A better model of the overall dynamics of the duplex caused by this lesion may lead to a more thorough understanding of these pathways. Investigating the effect due to variations of length and charge on these ICLs may lead to a better understanding of the recognition pathways involved in crosslink repair. More extensive knowledge of the recognition events may lead others to develop more effective chemotherapeutic agents.

Two questions will be addressed through these studies. First, the comparison of experimental and computational data of the stable analogs will validate the calculations. Second, the comparison of the simulation data of the stable analogs and native NM will validate the use of stable analogs in experiments as a substitute for the unstable NM.

1.3.2 Quantifying base eversion

Elaborate cellular repair processes have evolved to combat various types of DNA damage. One of these repair pathways is called base excision repair (BER).[98] In BER, individual glycosylases excise specific lesions that are generally more common than in other repair pathways, such as in nucleotide excision repair (NER).[8] The repair processes of BER are done with high fidelity and little or no energetic consumption. In general, all glycosylases share similar functionality: they evert the damaged base out of the duplex DNA for excision while inserting a wedge through the minor groove. Of particular interest in this work is studying the eversion pathway of 8-oxo-G when complexed to the Formamidopyrimidine DNA glycosylase (Fpg). To computationally study the pathway for eversion of 8-oxo-G, a metric to quantify eversion is needed, therefore a pseudodihedral angle was chosen. This eversion angle was shown to give substantial improvement over previous definitions.[43,47,99]

1.3.3 What role does Fpg's aromatic wedge play in damage recognition?

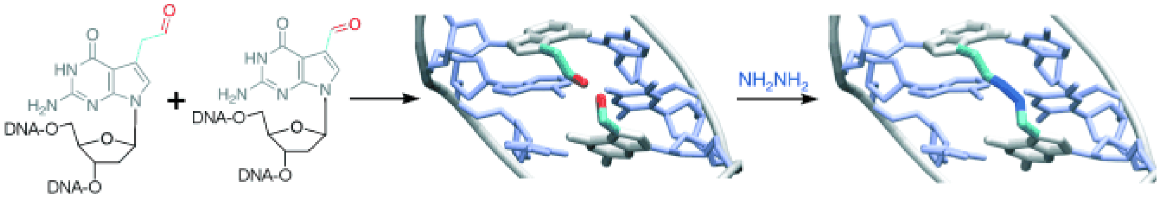
As mentioned earlier, BER glycosylases share structural traits when recognizing DNA damage. One particular functionality that BER glycosylases share is using aromatic or aliphatic residue(s) to probe DNA and destabilize the base pair of interest, facilitating eversion.[100] This behavior is seen with the bacterial glycosylase Fpg, in which a Phenylalanine (113) wedge is used to probe for damage. In this study, we are particularly interested in understanding the role that the Phe113 wedge plays in destabilizing the 8-oxo-G:C base pair to facilitate eversion. To accomplish this, we computationally mutated the Phe113 wedge to an alanine residue to better understand how the removal of the aromatic ring affects the overall profile to eversion. To measure the energetic profiles to eversion for 8-oxo-G and the F113A mutant, the partial nudged elastic band method was coupled with umbrella sampling.[91] After obtaining the energetic profiles to eversion, it was found that removing the wedge stabilizes the intrahelical state. Through the use of structural analysis and pairwise energy decomposition, we were able to uncover the key interactions that result from the presence of the intercalating wedge.

1.3.4 How does Fpg discriminate between a damaged 8-oxo-G and an undamaged G in DNA?

There are fewer than 12 known DNA glycosylases in Eukaryotic cells.[100] These glycosylases recognize and excise different types of damaged nucleobases in DNA with great efficiency. What is not known is exactly how these glycosylases distinguish damaged from undamaged DNA. In this study, the discriminatory behavior of the bacterial DNA glycosylase Fpg is explored when interrogating a G or 8-oxo-G nucleobase. Through the use of partial nudged elastic band theory and umbrella sampling, we investigated whether base eversion goes through the major groove or minor grooves. From these energetic pathways, a transiently stable exo-site in the major groove eversion path of Fpg was found where discrimination occurs between G and 8-oxo-G. As a result of two residues (Arg263/Gly264) that specifically recognize G in the exo-site, the eversion pathway constricts, thus no longer allowing G to continue on the eversion path. This constriction of the G eversion pathway results in an increase in the energetic barrier to eversion. Three specific interactions in the exo-site of Fpg recognize 8-oxo-G, the second 5' phosphate group and the conserved Asn173. These interactions with 8-oxo-G result in an opening of the pathway, creating a low-energy pathway to further eversion. By removing the key Asn173 interaction, the exo-site becomes unstable for 8-oxo-G as well as lowering the energetic barrier for further extrusion.

Chapter 2 Synthesis and molecular modeling of a nitrogen mustard DNA interstrand crosslink

Abstract



Over 60 years after nitrogen mustards (NMs) were the first agents used to treat tumors by chemotherapy, we provide a method to generate the main DNA adduct formed by NMs and validate them using molecular dynamics simulations (see figure above). We are able to provide amounts that permit extensive structural and biological studies. In our simulations we provide direct evidence that the analogs created give similar amounts of distortion as compared to the naturally occurring NM crosslink and have similar difference when compared directly to undamaged DNA.

Acknowledgments

The material presented in this chapter contains direct excerpts from the manuscript *Synthesis and Molecular Modeling of a Nitrogen Mustard DNA Interstrand Crosslink* by Angelo Guainazzi, Arthur J. Campbell, Todor Angelov, Carlos Simmerling and Orlando D. Schärer, published in "Chemistry A European Journal" in 2010, volume 16, pages 12100-12103. The manuscript was written by Angelo Guainazzi and Arthur J. Campbell with suggestions and revisions from Professor Carlos Simmerling and Orlando Schärer. Angelo Guainazzi and Todor Angelov conducted the experimental studies. Arthur J. Campbell performed the simulations and analysis.

2.1 *Introduction*

DNA interstrand crosslinks (ICLs) are formed by bifunctional agents with the ability to covalently link two strands of duplex DNA. ICLs are extremely cytotoxic, since they block essential processes such as DNA replication and transcription. Based on these properties, agents such as nitrogen mustards (NM) or cisplatin are widely used in cancer chemotherapy.[101] ICLs induce a number of biological responses that counteract the therapeutic effects of crosslinking agents. However, the elucidation of the mechanism by which these responses cause resistance in tumor cells has been hampered by the lack of efficient methods to generate ICLs formed by antitumor agents. Current approaches toward the synthesis of site-specific ICLs have yielded mostly models of the clinically relevant lesions or ICLs formed by endogenous agents.[102-111] A number of recent studies have shown that ICLs with different structures are processed in distinct ways (see recent reviews[112-114]), emphasizing the need to generate substrates containing the ICLs formed by the clinically important drugs.

2.2 *Methods*

2.2.1 Parameterization details

For the nitrogen mustard (**1**) and the analog (**2**) systems new parameters were generated for the crosslinked residues, in a manner previously undertaken in our laboratory.[115] The parameterization was accomplished in two steps: in step 1 the partial charges were generated by RESP fitting and in step 2, analogous atom type information was assigned from similar atom types in ff99SB with the parmbsc0 backbone additions and the GAFF forcefields to complete any unknown atom types for residues of (**1**) and (**2**).[78,116,117] The partial charges and atom types on the phosphates and sugars were kept consistent with the ff99SB and parmbs0 parameters for DNA.[78]

In stage 1, we created two guanine nucleotides in GaussView 3.0, bridging the two bases with the crosslink attached at the N7 position of each guanine. For (**2**) the N7 was replaced with a C7 for both residues (Figure 2-3).[118] Three conformations for each system were generated where the dihedrals of NM and (**2**) were rotated 60 degrees about the C11-C12-N13-C14 and C12-N13-N14-C15 dihedrals, respectively, to obtain three different structures for RESP fitting. These structures were then optimized in Gaussian 03 using the HF/6-31G* basis set.[118] RESP fitting was then used to combine the Gaussian electrostatic surface potentials of each conformation into single point charges for the bases and crosslinks in the new systems.[117,118] For (**2**), single point charges on the bases and crosslink were uniformly decreased after RESP fitting to make the net charge an integer (-2 for (**2**), which includes the two phosphate groups, sugars, bases and crosslink). For (**1**) an additional step was taken for generating the partial charges since the molecule is symmetrical. After the single point charges were generated for (**1**), the accompanying symmetrical atoms were then averaged. Next the single point partial charges of the (**1**) were uniformly increased to bring the net charge to an integer +1 for (**1**), the two crosslinked residues, this including the phosphate groups, sugars, bases and crosslink. In some cases it was necessary to change the actual name of atom type while keeping the parameters to address any overlap in atom type parameters (discussed further below).

In the parameterization of the two crosslinks, special care was given in the assignment of atom types. When defining new bonds, angles, dihedrals and improper dihedrals in the frcmod, the parameters defined in the original forcefield for the normal nucleotides used can be overwritten.[119] To avoid this problem, several new atom types (CH, CL, HL, NL and HM) were defined to prevent the normal definitions from being overwritten. The frcmod and prepin files for (**1**) and (**2**) are available from the authors upon request.

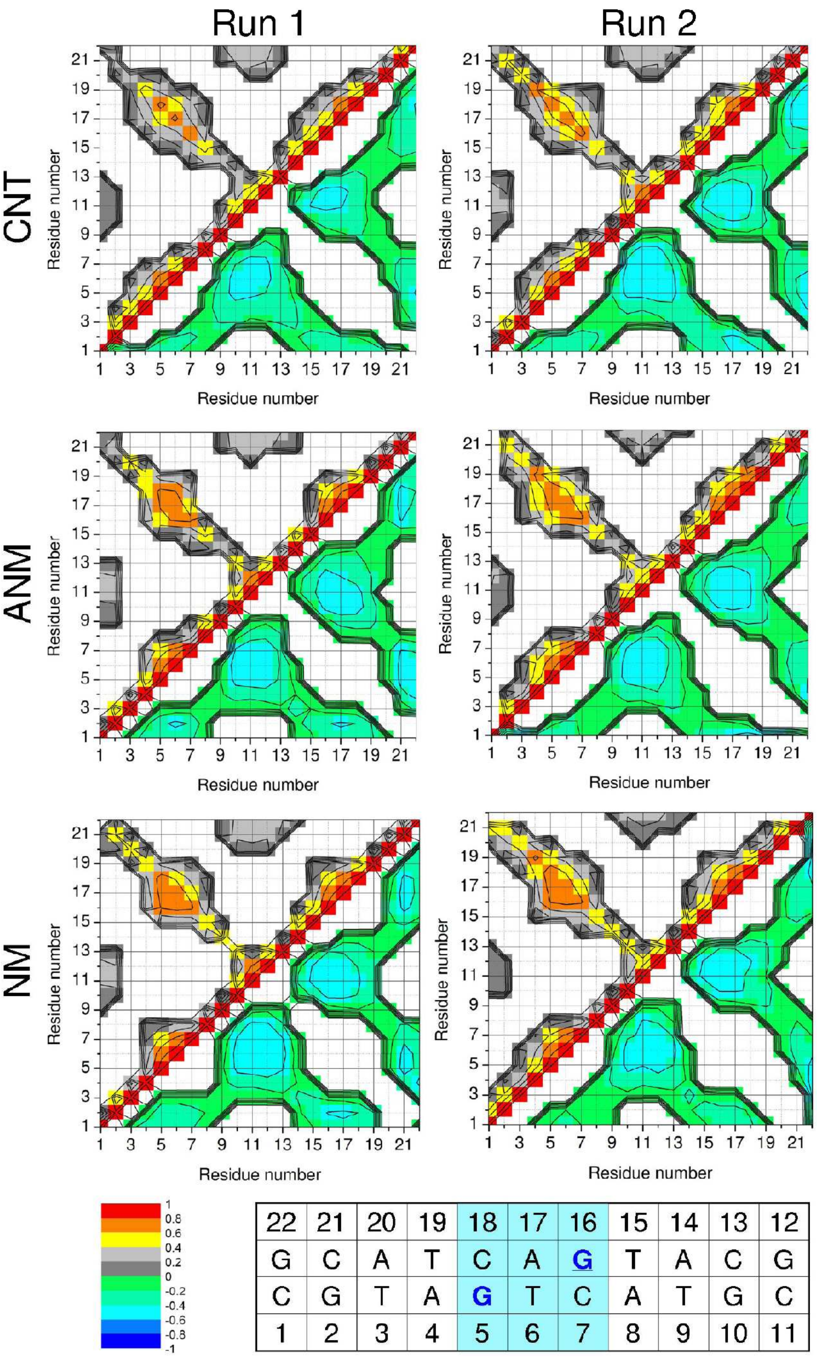
2.2.2 Simulation details

The Amber 9 suite of programs was used for all molecular mechanics calculations.[120] In this work three systems were simulated: the control (**C**), nitrogen mustard (**1**) and analog nitrogen mustard (**2**). The ff99SB forcefield with the parmbsc0 DNA backbone parameters was used for the majority of this work with additional GAFF parameters used on the crosslink to supplement regions that were not defined in the ff99SB forcefield.[119] The initial coordinates of each ICL system were generated in Nucgen using the identical sequence except for the crosslinked residues 5 and 16 (Figure 2-1). For each system, run-1 was built as A-form DNA and run-2 was built as B-form DNA to provide an independent measure of sampling convergence.

For the crosslinked systems, after the core sequence was built in Nucgen, the atom names of residues 5 and 16 were then changed to reflect the new parameter atom names of the crosslinked residues.[120] Any atoms that were in guanine but not in the (**1**) or (**2**) systems were deleted. These structures were then loaded in Tleap and any missing atoms were added. After the systems were built the connectivity for the crosslink was verified through visual inspection in Xleap and changed to accurately represent Figure 2-1.[120] The systems were then solvated with the TIP3P explicit water model in a truncated octahedron periodic box using an 8Å buffer containing ~9,000 water molecules.[86] After solvation the control was treated to a process that involved minimization, equilibration and then a production run. The two crosslinked systems were treated to two consecutive rounds of minimization, equilibration, and then unrestrained dynamics. The (**1**) and (**2**) systems were simulated in two stages since the initial structures may not have been compatible with the crosslink. To address this issue the (**1**) and (**2**) systems were created with incomplete crosslinks, where a single bond was missing on the alkyl linker, thereby giving two small separate lesions on residues 5 and 16 (see Figure 2-1 for sequence numbering). For the first stage of the two crosslinked systems (**1** and **2**) the C12 to N13 bond on the crosslink was purposefully omitted from the initial structures, having an incomplete crosslink. These structures were minimized, equilibrated, and then simulated in unrestrained dynamics for 10ns. The simulations of these incomplete structures were necessary to allow the duplex to accommodate the crosslink without any initial major distortions (data not shown). For the second round of the ICL system process the 10ns trajectory from round one was analyzed and the C12 to N13 distance was measured. From this measurement, one structure towards the end of the 10ns unrestrained dynamics trajectory was extracted from each of the A/B form ICL systems (**1** and **2**) that had a low C12 to N13 distance (~3Å). The solvent was removed from these structures and the C12 to N13 bond was created. These structures were then resolvated, minimized, equilibrated, and simulated in unrestrained dynamics for 50ns. Further details of these steps are given below.

The two control systems were minimized for 1,000 steps at constant volume with a restraint force of 500 kcal/mol-Å2, which was applied to all atoms of the duplex. After minimization, equilibration was performed under constant pressure, with a timestep of 2fs and with SHAKE employed to restrain the distances for bonds involving hydrogen.[120,121] The equilibration was executed in five stages, where the restraint force holding all atoms of the duplex decreased incrementally. At the first stage all atoms of the duplex were held with a restraint force of 10.0kcal/mol-Å2 and the temperature was slowly raised to 330K over 100ps. The temperature of 330K was chosen to reduce any energetic barriers in sampling. The next four stages of equilibration were then simulated at constant temperature with a restraint force of 5.0, 2.5, 1.0 and 0.5 kcal/mol Å2 respectively at 100ps increments. After equilibration the (**C**) system was simulated for 50ns of unrestrained dynamics at constant pressure, with a timestep of 2fs and SHAKE on.[121] Temperature was held constant with the weak coupling algorithm with a coupling constant of 1 ps.[122] A time step of 0.002 ps was used.

The nitrogen mustard system (**1**) and analog nitrogen mustard system (**2**) were simulated in two stages. As mentioned above, the first stage contained an incomplete crosslink (missing the C12 to N13 bond). The second stage contained the complete crosslink. In stage 1, the two system was minimized and equilibrated in the same manner as that described above for the control with the exception of the minimization which was minimized in two parts. The first part was 1,000 steps with a restraint force of 500 kcal/mol Å2 on all atoms of the duplex. The second minimization was 1,000 steps with a restraint force of 50 kcal/mol Å2 on all atoms. After stage 1 equilibration, the system was simulated in unrestrained dynamics for 10ns at constant temperature (330K). For the 10ns simulation, the C12 to N13 distances were measured and a structure was extracted for each run. As mentioned above the structures that were extracted had a low C12 to N13 distance (~3 Å) and were extracted from a time point that was near the end of the 10ns simulation in order to allow the system to accommodate the incomplete lesion. Next, the waters from these structures were removed in ptraj and were reloaded into Xleap where we created the final connectivity between the two atoms (making the C12 to N13 bond complete) to make the complete crosslinks.[120] In stage 2, after the crosslinks were completed as shown in the top of Figure 2-2 the systems were solvated, minimized, equilibrated and ran a final production simulation in the same manner as stage 1 but the production simulation was extended to 50ns. The final 50ns of unrestrained dynamics generated was used for our final analysis in the text (Figure 2-8) and supplemental (Figure 2-1 and Table 2-1).



22	21	20	19	18	17	16	15	14	13	12
G	C	A	T	C	A	**G**	T	A	C	G
C	G	T	A	**G**	T	C	A	T	G	C
1	2	3	4	5	6	7	8	9	10	11

Figure 2-1**.** Correlation coefficient matrix plots of runs 1 and 2 (left and right columns, respectively) of uncrosslinked control (**C**), ICL (**2**) and NM ICL (**1**) systems. The correlation coefficients were calculated using ptraj in the Amber package.[120] In these plots the correlation coefficient is the measure of each residue's movement relative to all others, as well as to itself.

Each residue represents the heavy atoms of its base group. The positive and negative correlation coefficients (n) are separated into two regions. The upper left triangular region of each plot represents the positive correlation coefficients ($n \geq 0$) and the lower right triangular regions represent the negative coefficients ($n<0$) of each plot. The bins on each plot represent the average correlation coefficient across each 50ns simulation for the two residues specified by the axes. For clarity, the respective residue numbers of the duplex sequence are shown at the bottom of the Figure. The sequence region highlighted in cyan represents the three base pairs that are highly involved in the crosslink, and the bases that are crosslinked in (**1**) and (**2**) are in blue font. From the correlation plots it is apparent that the positive correlation coefficient region where the crosslink was introduced (highlighted in blue in the sequence chart on the bottom of the figure) increases in correlation for (**1**) and (**2**) when both are compared to the control. This observation implies that the introduction of an interstrand crosslink of the length and chemical makeup as observed in Figure 2-3 directly increases the amount of tandem movement of the residues around the crosslink.

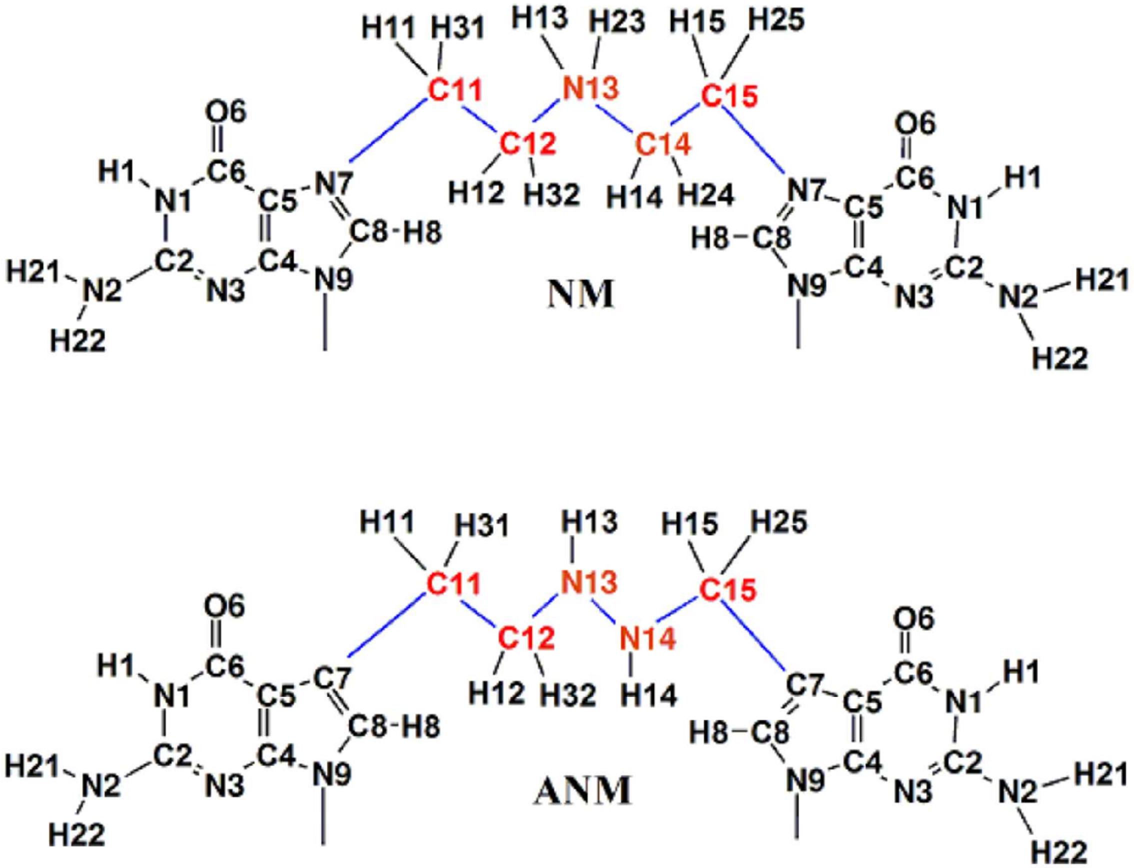


Figure 2-2. Atom names of (**1,** top) and (**2,** bottom). The heavy atoms of the crosslink are highlighted in red and their interconnecting bonds are in blue for clarity. These names match the names in the library files, which are provided as supplemental material.

Table 2-1. CURVES analysis measurement of the 50ns simulations for base pairs 5:18 and 7:16 of the **1**, **2** and **C** systems.[123] Corresponding residue numbers in the duplex can be seen at the bottom of Figure 2-1. From this data the 7:16 base pair buckling became more negative when a crosslink was introduced, while there is a buckle of -3.5 degrees for the control. For the analysis of the 5:18 base pair it was found that the propeller twist of the two crosslinked systems deviated most from the control rather than the buckling. The main conclusion that could be drawn from these measurements is that interstrand crosslinks gave some degree of local distortion but the type of resulting distortion was dependent on other factors besides the crosslink introduced. We speculate that the local distortions such as buckling and propeller twist from the crosslinks in this work (Figure 2-3) are sequence dependent, and the positioning of a purine/pyrimidine between the 1,3 ICL (at residue 6 or 17, see Figure 2-1) may have a significant effect on the amount of distortion. The precision was calculated from the two separate simulations from each system.

CURVES analysis				
System	Base pair	Measurement	Average (degrees)	Uncertainty (degrees)
2	5:18	Buckle	1.9	±0.0
2	7:16	Buckle	-21.8	±0.6
2	5:18	Propeller twist	-26.9	±0.8
2	7:16	Propeller twist	-12.7	±0.6
C	5:18	Buckle	-1.1	±0.1
C	7:16	Buckle	-3.5	±0.2
C	5:18	Propeller twist	-13.2	±0.1
C	7:16	Propeller twist	-8.3	±0.0
1	5:18	Buckle	-2.5	±0.3
1	7:16	Buckle	-13.2	±0.1
1	5:18	Propeller twist	-22.3	±0.2
1	7:16	Propeller twist	-18.3	±0.1

2.3 *Results and Discussion*

2.3.1 Experimental work

Here we report the synthesis of a stable NM ICL isostere (**2**) and use atomic detail simulations to show that it recaptures the structural and dynamic properties of the native NM ICL (**1**) (Figure 2-3.). To obtain the NM ICL isostere we used a strategy that we recently developed for the synthesis of major groove ICLs.[57]

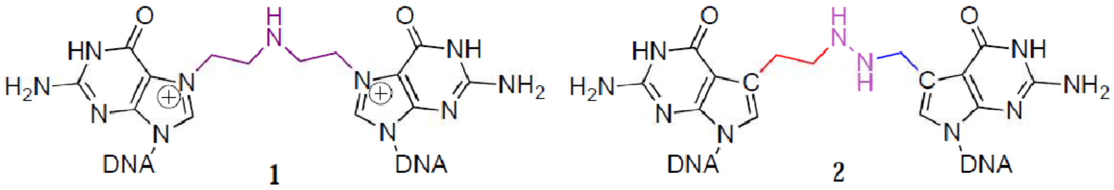


Figure 2-3. Structure of a NM ICL (**1**) an the stable analog (**2**).

This approach involves a double reductive amination reaction of two acetaldehyde functionalities (**5**) linked to the 7-position of G residues on complementary strands of dsDNA. One dG residue was substituted with 7-deaza-2′-deoxyguanosine to counteract the inherent lability of the glycosidic bond in N7-alkylated guanines.[124] This approach allowed the generation of a six-atom ICL by coupling two acetaldehyde groups with hydrazine. However, we were unsuccessful in generating ICL (**3**) with the five-atom bridge found in NM ICLs[61,125,126] (Figure 2-4), possibly because the reductive amination with ammonia was not powerful enough to introduce the strain in the DNA caused by the ~7.5 Å bridge of the NM.[57,127,128]

We reasoned that an ICL isosteric to those formed by NM might be formed by reaction of hydrazine with an acetaldehyde (**5**) and a formyl aldehyde derivative (**4**) of deazaguanine, exploiting the higher reactivity of hydrazine over ammonia (Figure 2-4).[57]

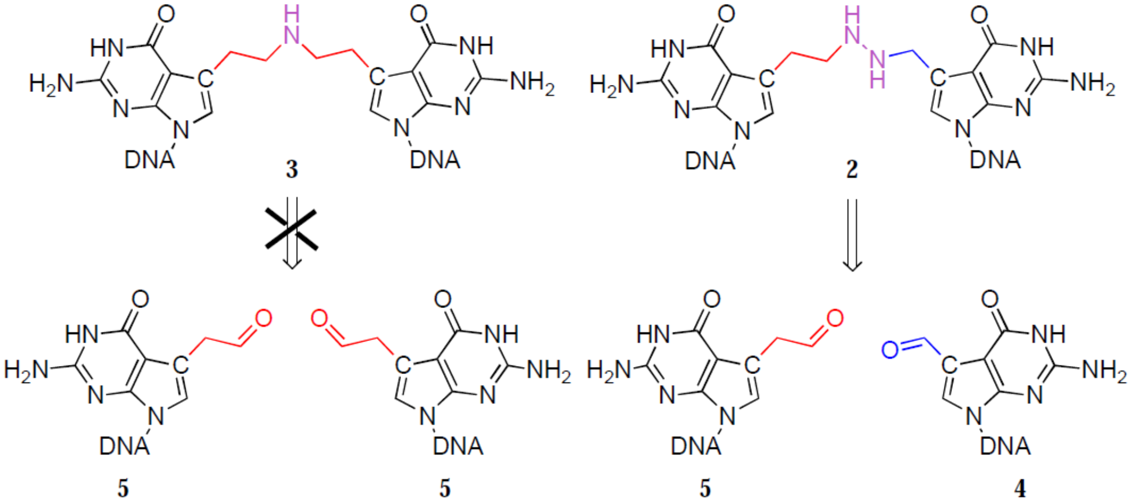


Figure 2-4. Formation of NM ICLs by reductive amination was not successful using two acetaldehyde precursors (**5**) and NH4Cl, prompting us to explore the generation of isostere (**2**) by linking precursors (**4**) and (**5**) with hydrazine.

We synthesized formyl aldehyde (**4**) in an analogous fashion to the previously reported synthesis of acetaldehyde (**5**), masking the aldehyde as a protected diol during solid-phase DNA synthesis.[57,129] The synthesis started with vinylation of a protected 7-iodo-7-deazaguanine derivative (**6**) using a Stille coupling reaction (Figure 2-5) to give compound (**7**). Oxidation of the allyl group to the diol (**8**), protection, and functionalization yielded phosphoramidite (**9**). The aldehyde precursors were incorporated into complimentary 20-mer oligonucleotides in a 5′-d(GNC) sequence (the preferred sequence context for NM ICL formation[61,125,126]) by solid-phase synthesis, and the oligonucleotides were deprotected and annealed.

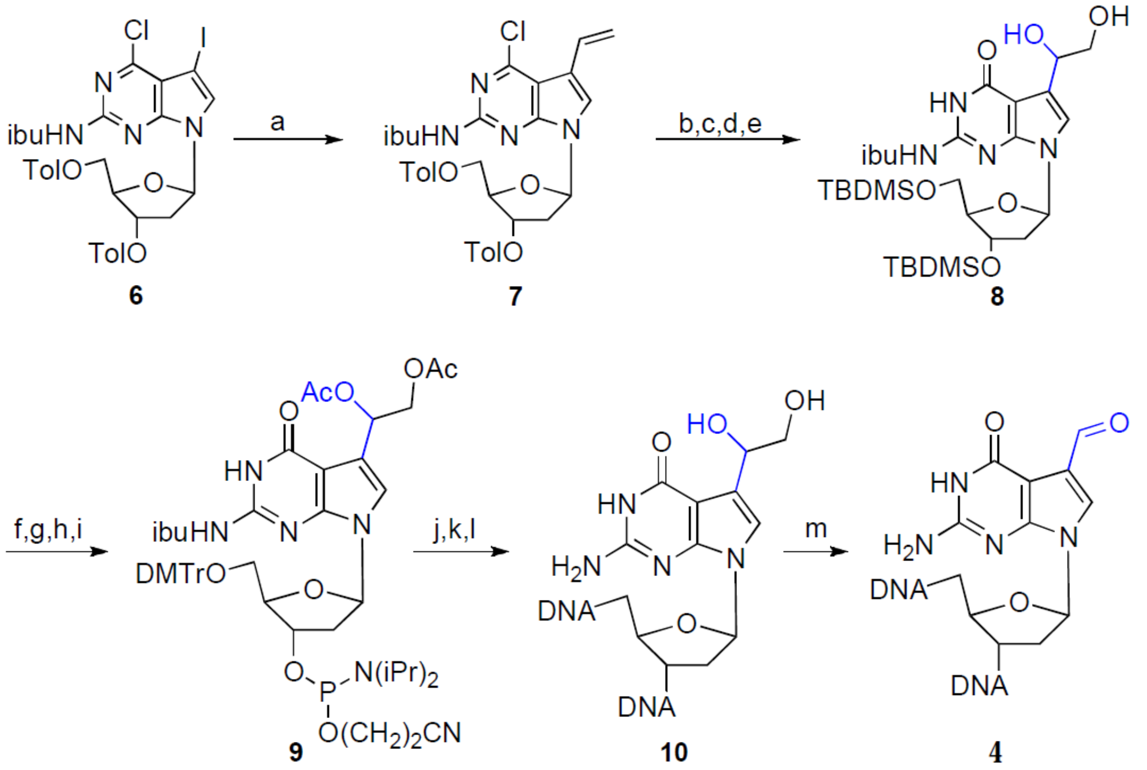


Figure 2-5. Synthesis of the formyl aldehyde 4: a) vinyl-Sn(Bu)3, Pd[P(Ph)3]4, toluene, 90 °C, 43 h, 70 %; b) pyridine-2-carboxaldoxime, N,N,N′,N′-tetramethylguanidine, dioxane, DMF, RT, 42 h, 88 %; c) NaOMe, THF, RT, 5 h, 97 %; d) TBDMS-Cl, imidazole, DMF, RT, 16 h, 93 %; e) OsO4, NMMO, THF, 0 °C, 3.5 h, 63 %; f) Ac2O, pyridine, RT, 1 h, 82 %; g) TBAF, AcOH, THF, RT, 20 h, 79 %; h) DMTr-Cl, pyridine, RT, 1 h, 74 %; i) iPr2NP(Cl)OC2H4CN, DIEA, CH2Cl2, RT, 1 h, 72 %; j) oligonucleotide synthesis; k) 33 % NH4OH, 50 °C, 12 h; l) annealing; m) NaIO4, RT, 12 h. TBDMS-Cl=tert-butyldimethylsilyl chloride; NMMO=N-methylmorpholine oxide; TBAF=tetra-n-butylammonium fluoride; DMTr-Cl=4,4′-dimethoxytrityl chloride; DIEA=N,N-diisopropylethylamine; ibu=isobtuyryl.

Following oxidation of the diols using periodate, the two aldehydes were coupled by double reductive amination with hydrazine and NaBH3CN.[57] The reaction along with appropriate controls was analyzed by denaturing PAGE.

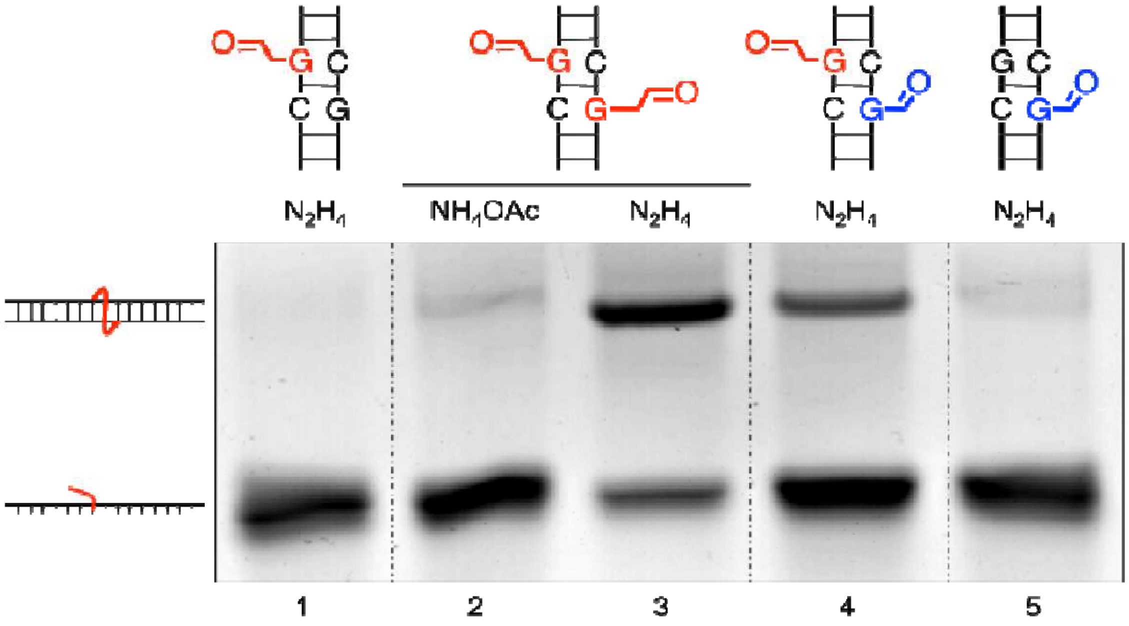


Figure 2-6. Analysis of ICL formation by denaturing PAGE with methylene blue staining. The duplexes, amine used and position of single-stranded or ICL-containing DNA are indicated. The sequences used were 5′-d(GTCACTGGTAXACAGCATTG) and 5′-d(CAATGCTXTCTACCAGTGAC) where X represents the modified G. The small amount of bands running as duplexes in lanes 2 and 5 are likely due to residual amount of duplex that was not denatured during electrophoresis.

As already discussed[57], ICL formation from two acetaldehyde precursors was successful with hydrazine, but not with ammonium acetate (Figure 2-4, lanes 2 and 3). However, reaction of a duplex containing **4** and **5** with hydrazine led to formation of the desired five-atom ICL **2**, evidenced by a band with the same mobility as the previously analyzed crosslink (Figure 2-6, lanes 3 and 4). Although the yield of the five-atom ICL was lower than that of the six-atom ICL

(~25 % vs. ~75 %), we were able to isolate the product by gel purification and electroelution in amounts exceeding 100 nmol. For simplicity of analysis we also synthesized ICL **2** in an 11-mer duplex and unambiguously identified it as the desired product by ESI-MS (*m*/*z* calcd: 6741.23, found: 6741.6, Figure 2-7 in the Supporting Information). If either aldehyde precursor was present only on one strand of the duplex, no significant amounts of the slower migrating species were observed, further demonstrating the specificity of ICL formation (Figure 2-6, lanes 1 and 5).

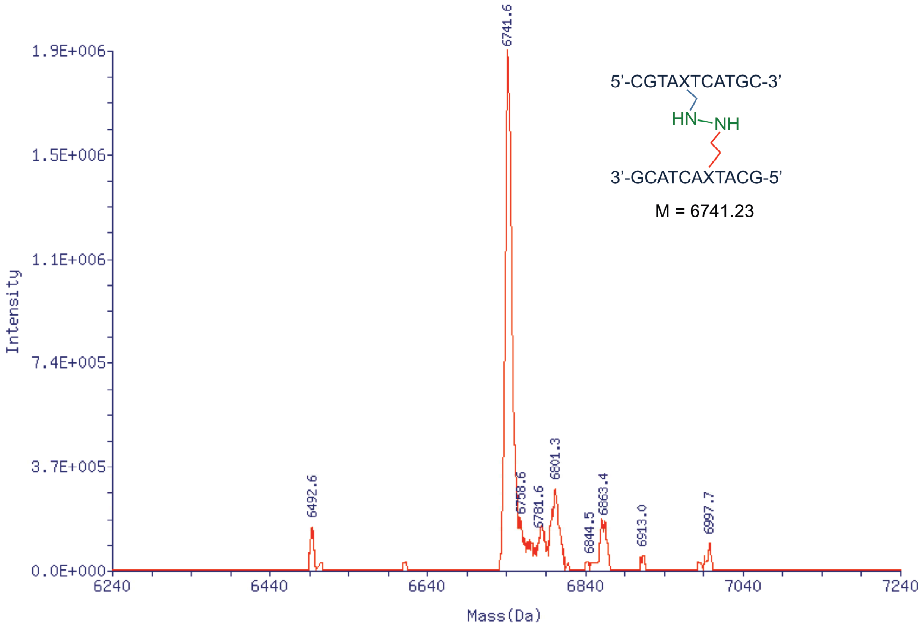


Figure 2-7. Mass spectrum of the crosslinked oligonucleotide (**2**). The structure and the calculated mass are indicated.

Molecular modeling was used to validate ICL (**2**) as a model for NM (**1**) and to compare the structural consequences of the two ICLs with the uncrosslinked control (**C**) in identical 11-mer sequences. We used the Amber simulation package with atomic detail and explicit water (see Supporting Information) to validate that (**2**) gives similar amounts of distortion when compared to the NM ICL (**1**) and consistent differences when both are compared to (**C**).[27] Two independent simulations of 50 ns were run for each of the (**1**), (**2**), and (**C**) systems. The duplex was stable throughout all six simulations.

The N7 to N7 distances of the NM ICL (**1**) overlapped well with the C7 to C7 distances of (**2**); both are restricted as compared to (**C**), which samples a broader range (

Figure 2-8, left).

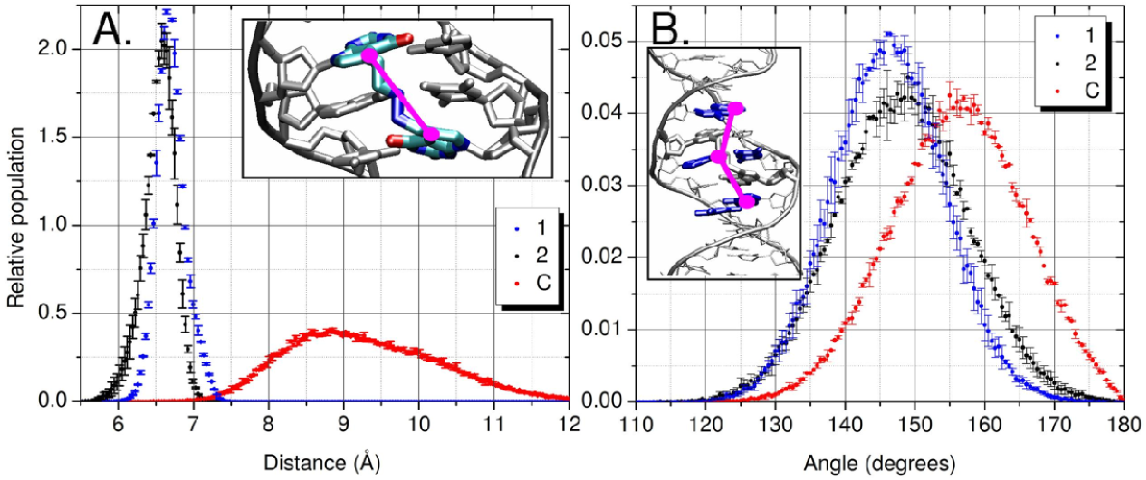


Figure 2-8. Data from MD simulations of **2** (black), **1** (blue) and **C** (red). Left: Distance measurement marked in pink, C/N7 to C/N7 atoms between residues 5 and 16. Right: Center of mass angle measurement, in which each point is defined by the heavy atoms of the base pair (highlighted in blue in inset picture): point 1 represents 4 A and 19 T, point 2 represents 6 T and 17 A and point 3 represents 8 A and 15 T.[77]

The decreased distance between the two crosslinked bases has a direct effect on the local distortion around the ICL, and is also influenced by the sequence (Figure 2-8 right, and Table 2-1). One measure of distortion is to examine the buckling and propeller twist[123] of the two crosslinked base pairs (Table 2-1 and Figure 2-1 for sequence). Since molecular mechanics force fields have limited accuracy, it is more important to note the trends rather than the specific values. It is well accepted that differences are more reliable in simulations than absolute values, due to cancellation of systematic error. The crosslinked guanine 16 that neighbors the large purine ring of adenine (residue 17) accommodates the crosslink by buckling of the 7:16 base pair (**C**: −3.5±0.2°, **1**: −13.2±0.1°, and **2**: −21.8±0.6°). The crosslinked guanine 5 that neighbors the small pyrimidine ring of thymine 6 has more room to move and accommodates the ICL by an increased propeller twist of the 5:18 base pair (**C**: −13.2±0.1°, **1**: −22.3±0.2°, and **2**: −26.9±0.8°). The addition of a covalent bridge between both strands of DNA nearly doubles residue to residue correlation in the central region of the duplex. Residues 5 and 16 have a correlation coefficient (*r*) of 0.29±0.04, 0.62±0.01, and 0.58±0.02 for (**C**), (**1**), and (**2**) respectively (Figure 2-1). ICLs (**1**) and (**2**) have a substantial decrease in flexibility (Figure 2-8, left) when compared to the uncrosslinked reference CNT, resulting in the formation of a slight kink in the duplex of both (**1**) and (**2**) to accommodate the ICL (Figure 2-8, right).

To test for local duplex bending at the crosslink site we measured an angle that would encompass the tightening of the 7- to 7-position of residues 5 and 16 (Figure 2-8, right). The smaller angles sampled by (**1**) and (**2**) suggest that the analogue slightly bends the duplex outside the crosslink in the same manner that the NM ICL simulations do (Figure 2-8, left). The slight bending of the central region of the ICL simulations of (**1**) and (**2**) compared to (**C**) is qualitatively similar to past experimental work.[127,128]

To measure the local differences between the three systems the glycodic torsion angle and the sugar pucker were measured at each picosecond throughout each 50ns simulation. From this data we were able to measure the free energy landscapes based on these populations of these two reaction coordinates (Figure 2-9, Figure 2-10 and Figure 2-11). It is evident by comparing these three figures that the energetic landscapes of the Analog and Nitrogen mustard have similar landscapes in residues 6, 16 and 17 while these are different for the control system. These differences are most likely caused from the introduction of the crosslink in the duplex which restricts local movements. The most populated structures for run 1 of the entire 50ns for the control, analog and nitrogen mustard were plotted (Figure 2-12). These structures have few differences since they are static structures and the true differences only surface when a figure is represented by the whole trajectory.

Circular dichroism (CD) spectra were recorded to gain experimental insight into to what extent the NM ICLs containing oligonucleotide (**2**) deviates from B-form DNA. The CD spectra of (**2**) and an unmodified duplex of the same sequence displayed the characteristic features of B-form DNA (Figure 2-13 in the Supporting Information), consistent with our molecular dynamics simulations, indicating the NM ICL induces only a minor bend in the DNA. It has been shown that only more dramatic distortions, such as the ones induced by cisplatin ICL result in significant changes in CD spectra.[111,130]

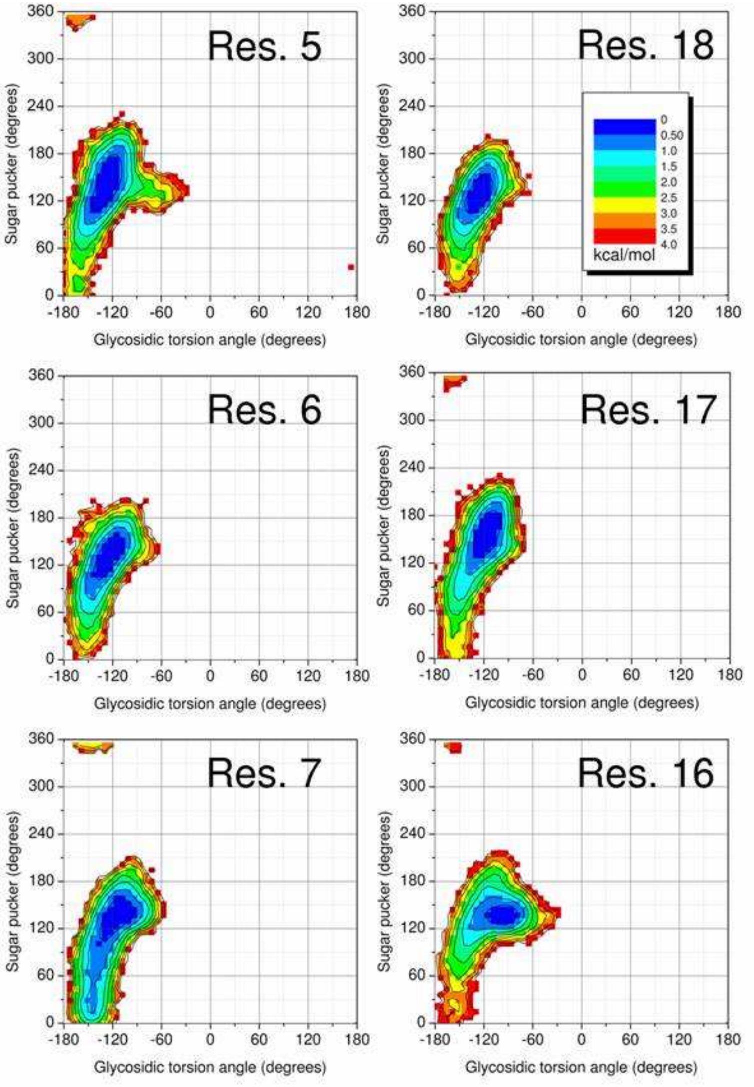


Figure 2-9. Control system energetic landscape of sugar pucker (degrees) vs. glycosidic torsion angle (degrees) for control system duplex. The central three basepairs are plotted (residues 5-7 and 16-18). Energy is measured in kcal/mol.

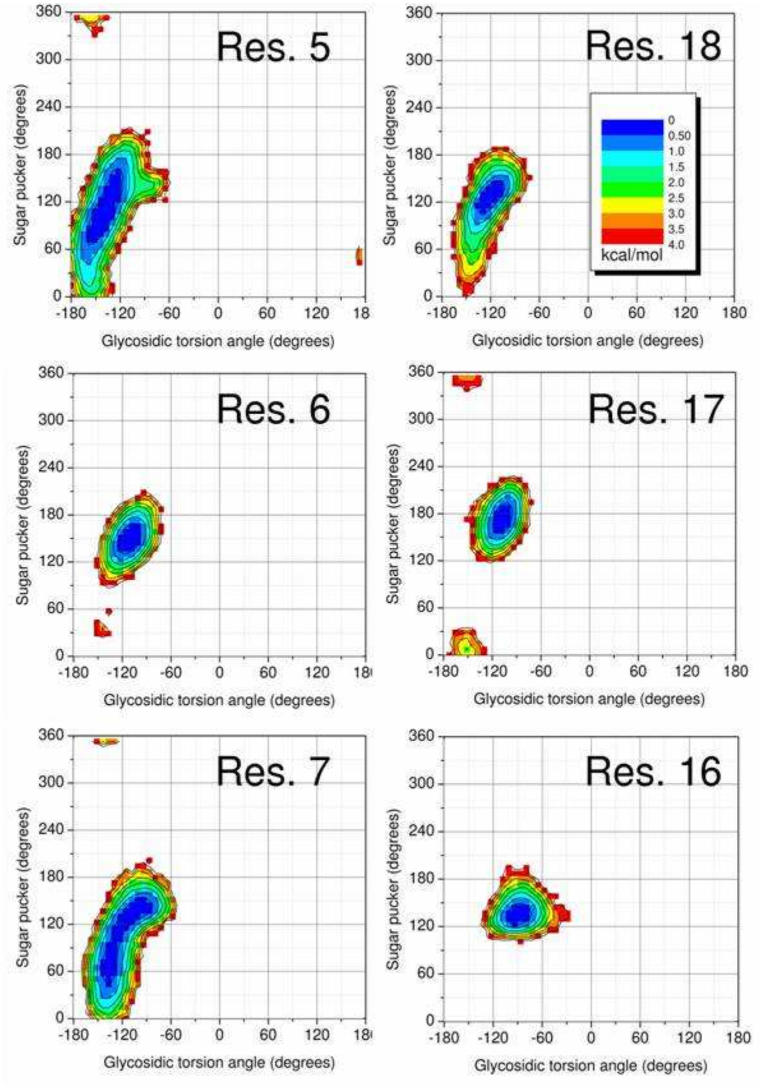


Figure 2-10. Nitrogen mustard system energetic landscape of sugar pucker (degrees) vs. glycosidic torsion angle (degrees) for nitrogen mustard duplex. The central three basepairs are plotted (residues 5-7 and 16-18). Energy is measured in kcal/mol.

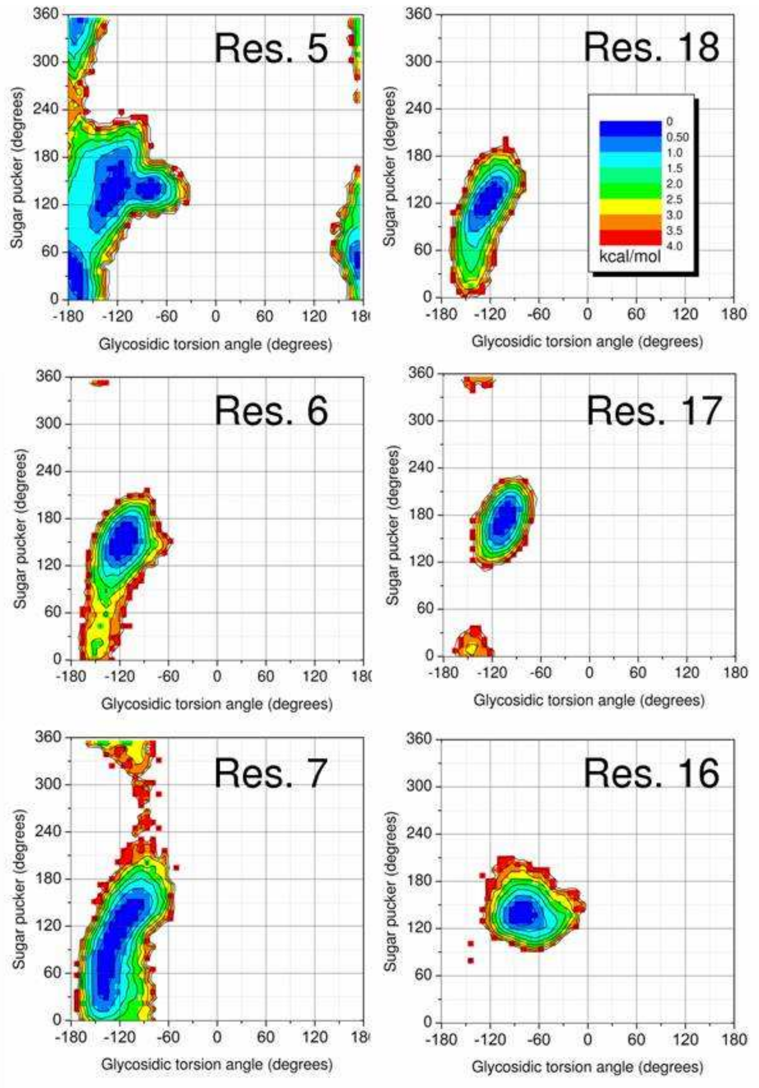


Figure 2-11. Stable analog system energetic landscape of sugar pucker (degrees) vs. glycosidic torsion angle (degrees) for stable analog duplex. The central three basepairs are plotted (residues 5-7 and 16-18). Energy is measured in kcal/mol.

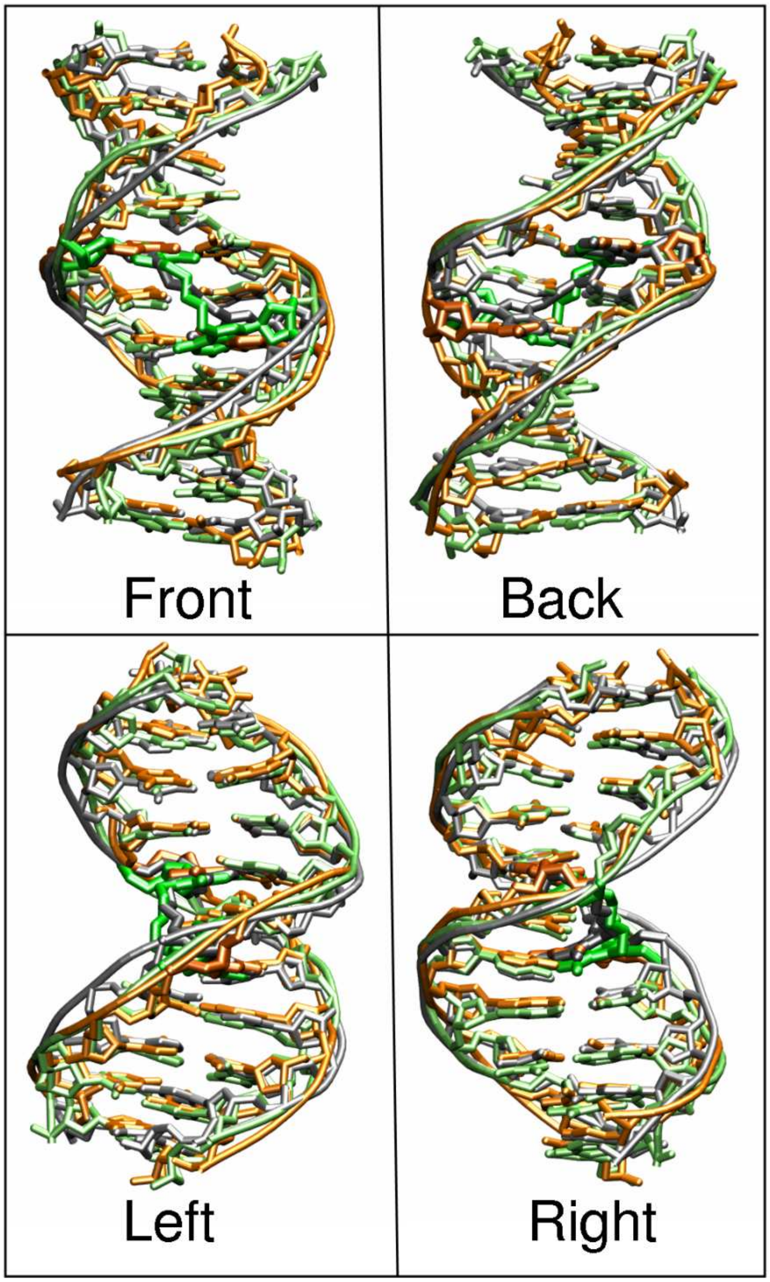


Figure 2-12. The most populated structures for run 1 of the entire 50ns of the control (orange), analog system (green) and nitrogen mustard (gray) overlapped on their respective phosphorus's.

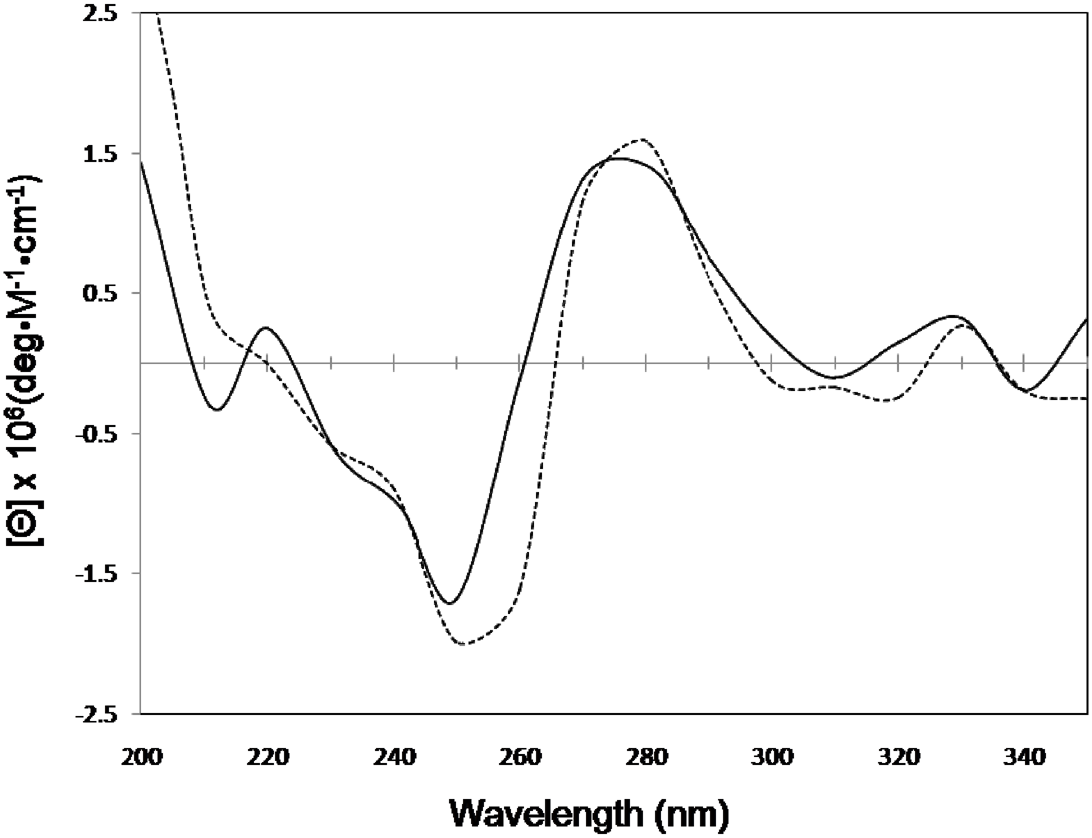


Figure 2-13. CD spectra of normal undamaged DNA (solid line) and DNA with a 1,3 interstrand crosslink analog (dashed line).

2.3.2 Experimental section on crosslink formation

A solution of the single-strand oligonucleotides (25 nmol) in NaCl (125 µL, 10 mM) was heated to 95 °C and allowed to cool to room temperature over a period of 4 h to allow for duplex formation. After addition of sodium phosphate buffer (15 µL, 1 M, pH 5.4) and $NaIO_4$ (10 µL, 50 mM) the reaction mixture was kept in the dark overnight at 4 °C. Excess $NaIO_4$ was removed by centrifugation through Microcon columns with a 3 K cutoff (Millipore). Then aqueous hydrazine (10 µL, 5 mM) and $NaCNBH_3$ (10 µL, 0.5 M) were added and the reaction mixture was left overnight at room temperature in the dark. ICL formation was assessed by electrophoresis on a denaturing 20 % polyacrylamide gel. The band containing the crosslinked oligonucleotide was excised from the gel and the DNA was extracted by electroelution using D-Tube™ Dialyzer (Novagen) or the Elutrap™ (Schleicher & Schuell) device. Full experimental details and computational methods are available in the methods section above.

2.4 *Conclusions*

In summary, we describe the synthesis of the stable NM ICL analogue (**2**) by post-synthetic double reductive amination. Although this ICL has three atoms substituted with respect to the NM ICL (**1**) (the two N at the 7-positions of dG by C and one C by N in the hydrazine-formed bridge), our molecular dynamics simulations show that the two ICLs affect DNA structure and motion in equivalent ways. The availability of stable, site-specific NM ICLs will enable studies of the structural consequences and biological responses induced by NM ICLs,[131,132] more than sixty years after NMs were the first agents to be used in cancer chemotherapy.[133] Such studies will provide new insights into how tumors become resistant to treatment by crosslinking agents.

Chapter 3 An improved reaction coordinate for nucleic acid base flipping studies

Abstract

Base flipping is a common strategy utilized by many enzymes to gain access to the functional groups of nucleic acid bases in duplex DNA which are otherwise protected by the DNA backbone and hydrogen bonding with their partner bases. Several X-ray crystallography studies have revealed flipped conformations of nucleotides bound to enzymes. However, little is known about the base-flipping process itself, even less about the role of the enzymes. Computational studies have used umbrella sampling to elicit the free energy profile of the base-flipping process using a pseudodihedral angle to represent the reaction coordinate. In this study, we have used an unrestrained trajectory in which a flipped base spontaneously reinserted into the helix in order to evaluate and improve the previously defined pseudodihedral angle. Our modified pseudodihedral angles use a new atom selection to improve the numerical stability of the restraints and also provide better correlation to the extent of flipping observed in simulations. Furthermore, on the basis of the comparison of potential of mean force (PMF) generated using different reaction coordinates, we observed that the shape of a flipping PMF profile is strongly dependent on the definition of the reaction coordinate, even for the same data set.

Acknowledgments

The material presented in this chapter contains direct excerpts from the manuscript by Kun Song, Arthur J. Campbell, Christina Bergonzo, Carlos de los Santos, Arthur P. Grollman, and Carlos Simmerling, published in the "Journal of Chemical Theory and Computation" in 2009, volume 5, pages 3105-3113. The manuscript was written by Kun Song and Arthur J. Campbell with suggestions and revisions from Professor Carlos Simmerling. Dr. Kun Song and Arthur J. Campbell performed the simulations and data analysis.

3.1 *Introduction*

Base flipping (also known as base eversion) is a type of local DNA motion in which a base group loses the hydrogen bonds with its base pair partner and is everted from the intra-helical to extra-helical position.[134] Base flipping was first observed in the DNA/methyltransferase complex X-ray crystal structure.[135,136] Studies have shown that base flipping is a common strategy for enzymes to read and chemically modify base groups which are otherwise protected by their base pair partner, or their own sugar and phosphate groups.[134] A variety of these enzymes exist, such as methyltransferases, glycosylases, and endonucleases. A number of crystal structures with everted DNA base groups inside the active site of the enzyme have been published.[66,70,137-139] These structures reveal the conformations of the everted base groups, but they provide little insight into the conformational changes that occur during the flipping process and, more importantly, the possible transient role of enzyme functional groups in the facilitation of base flipping. The rate of base flipping can be measured experimentally by methods such as proton exchange.[140,141] However, studies have shown that this method may overestimate the flipping rate since proton exchange may occur in structures with limited solvent accessibility and thus not require complete base flipping.[43]

Several computational approaches have been applied to this subject.[43,45-51,142] It is currently necessary to force eversion using restraints in order to model the process during computationally tractable simulations since uncatalyzed base flipping occurs on the millisecond time scale.[140] In one of the earliest studies, Keepers et al. used a distance restraint between the N1(pyrimidine) and N3(purine) to force the base pair to break.[54,55] However, the distance restraint cannot specify which of the two bases to flip, or distinguish between major or minor groove flipping pathways. Inspired by the correlation between the base opening angle and the ζ

dihedral angle seen in the crystal structure, Chen et al. applied restraint forces on the ζ angle and glycosidic angles of the target nucleotide to force base pair opening.[53] This procedure assumes that only two backbone dihedral angles of the targeted nucleotide are responsible for base flipping, which may not be generally true. It has been shown that using this method can generate artificial conformations.[49] More promising approaches have employed more sophisticated reaction coordinates; these have been applied by Lavery and co-workers using internal coordinates[45,46,51,142] and MacKerell and co-workers using a center of mass pseudodihedral angle[43,47,48,50] (the latter is hereafter referred to as the CPD angle). The results from these two approaches are in reasonable agreement. Due to the nature of its definition using a standard dihedral angle with points defined by center of mass groups, MacKerell's pseudodihedral angle method is relatively easier to implement in current molecular dynamics simulation algorithms. An excellent recent application of this method to cytosine 5-methyltransferase from HhaI suggests that the enzyme shifts the equilibrium to the flipped state by destabilizing the DNA duplex and stabilizing the everted conformation.[47]

In the present study, we first tested the CPD definition to study base flipping using the Amber simulation package and then improved the CPD definition.[27] Our results show that MacKerell's highly valuable CPD definition can have several potential disadvantages in spite of its multiple strengths. Using the traditional CPD definition, we find that there are large energy fluctuations and occasional simulation instability when the base is everted. Another disadvantage is that there is not a unique mapping between the CPD angle and the extent of eversion; we observed that significantly different structures can have comparable CPD angles when using the previous definition. On the basis of our molecular dynamics (MD) simulations of spontaneous base pair formation in unrestrained simulation, we identified several reasons for the weaknesses described above. We thus modified the CPD definition to employ two separate definitions of different groups of atoms to define the pseudodihedral angle (hereafter referred to as CPDa and CPDb); we find that these better represent base flipping. Simulations using the CPDa/b reaction coordinates were able to give a more reliable representation of the base-flipping pathway with improved correlation between the reaction coordinate and the extent of eversion, along with improved simulation stability. The two new CPDa/b angles were applied to calculate the free energy profile of base flipping for guanine in a DNA duplex. The free energy barrier of the base flipping is comparable with that of previous studies. The relationship between the specific definition of the reaction coordinate and the overall shape of the resulting free energy profile is discussed.

3.2 *Methods*

3.2.1 System preparation

The initial structure was standard B-form duplex DNA built using the NUCGEN program in the Amber simulation package.[27] The DNA sequence used is shown in Table 3-2 below. The targeted central base pair for flipping was C10:G24, and G24 was the base group (hereafter referred to as the target base group) for base flipping.

Table 3-2. Sequence of the duplex DNA used in our simulations[a]

	1	2	3	4	5	6	7	8	9	**10**	11	12	13	14	15	16		
5’	A	G	G	T	A	G	A	T	C	**C**	G	G	A	C	G	C		
		C	C	A	T	C	T	A	G	**G**	C	C	T	G	C	G	T	5’
		32	31	30	29	28	27	26	25	**24**	23	22	21	20	19	18	17	

a. The base pair C10:G24 is the targeted central base pair, and G24 is the target base for flipping.

The initial coordinate and parameter files were created using the LEAP module of Amber on the basis of the structure generated by NUCGEN, solvated in truncated octahedron boxes with a minimum 8 Å buffer between the box edge and the nearest solute atom. The TIP3P water model[86] was used to explicitly represent water molecules. The DNA parameters employed ff99,[79,116] with the parmB-SC0 modified alpha/gamma torsional terms.[81] These coordinates were used for all simulations.

3.2.2 Molecular dynamics simulations

All molecular dynamics simulations were carried out with the SANDER module in Amber.[27] Following the procedure used in previous studies,[63,143] the solvated systems were minimized and equilibrated in three steps: (i) 50 ps MD simulation[144] with DNA atoms constrained and movement allowed only for water; (ii) five 1000-step cycles of minimization, in which the positional restraints on the DNA were gradually decreased; (iii) four cycles of 5000-step MD simulation with decreasing restraints on the DNA. A final 500 ps of MD was performed without restraints. The resulting structures were used in umbrella sampling simulations.

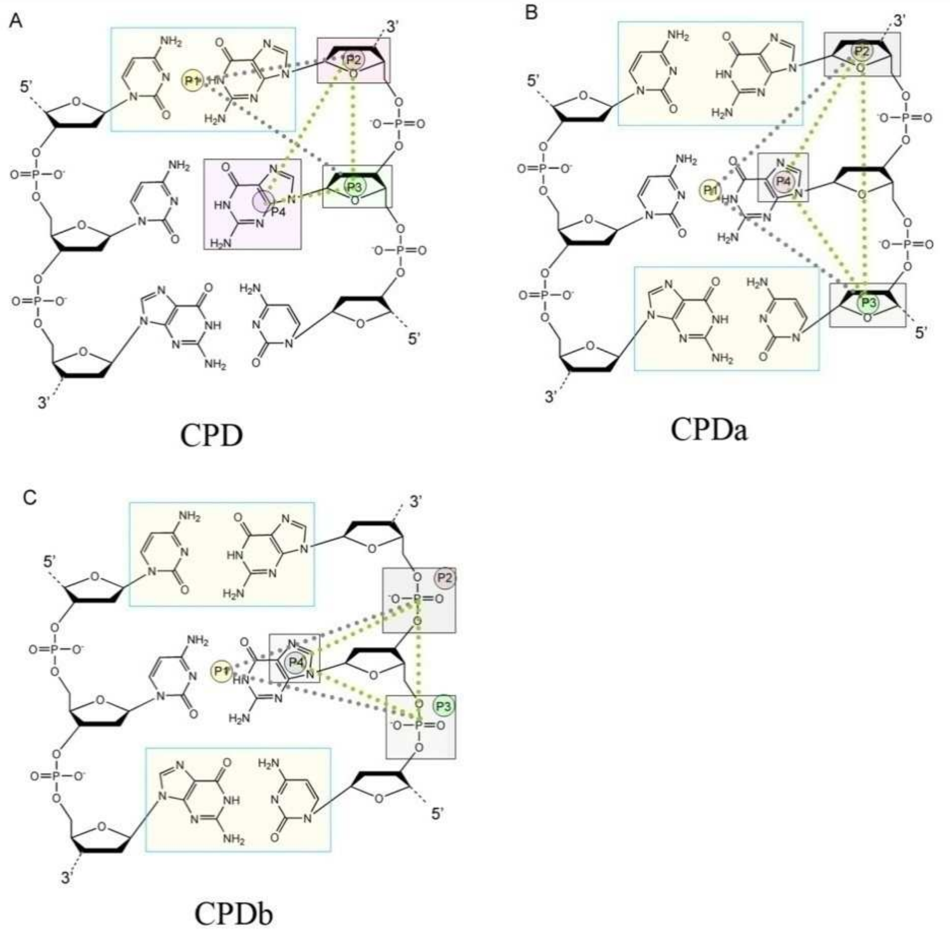
SHAKE[121] was used to constrain bonds involving hydrogen atoms. The nonbonded cutoff was 8 Å. The particle mesh Ewald method[87,88] was used to calculate long-range electrostatics. Constant pressure (1 atm) and temperature (330 K, slightly elevated to improve sampling) were maintained by the weak coupling algorithm with a coupling constant of 1 ps.[122] A 0.002 ps time step was used.

3.2.3 Structural analysis

To obtain unbiased structural analyses on the three types of reaction coordinates presented in this work, we generated an unrestrained MD simulation starting from a structure containing an everted target base, which then spontaneously reinserted into the duplex during the simulation. In this simulation, the same sequence of DNA duplex and the simulation conditions described in the previous sections were used. The root-mean-square deviation (rmsd) of all heavy atoms in the central three base pairs was calculated, with the reference structure of a standard B-form duplex DNA. The glycosidic angle of the G24 nucleotide was also measured, using the dihedral angle O4′−C1′−N9−C4. The distance between the central base pair units is represented by the distance between N1 of G24 and N3 of C10.

3.2.4 The definition of the base-opening dihedral angle

Umbrella sampling[94-96,145] was used to calculate the potential of mean force (PMF) as a function of our new center of mass pseudodihedral angles CPDa/b. The definitions of two proposed variations on this new flipping metric, CPDa and CPDb, are shown in Figure 3-1. The Sander module of Amber9 was modified to support these restraints. The four points for each dihedral form two triangular planes which share one side defined between P2 and P3. The base opening angle is defined by the angle between these two planes. Changes from the definition of MacKerell at al. involve the use of both flanking base pairs for the P1 center of mass (as opposed to only one flanking pair) and using either the sugars or phosphates flanking the flipping base as points P2 and P3. Point P4 was defined using only the five-member ring of the purine in order to remove the influence of glycosidic rotation on the flipping angle, which occurs if the entire purine base is included in COM point P4 (data not shown).



CPD

CPDa

CPDb

Figure 3-1. Definitions of various reaction coordinates for base eversion. (**A**) CPD: MacKerell et al.'s original COM pseudodihedral angle definition. (**B**) CPDa: the modified COM pseudodihedral (CPD) angle definition, in which p1 is defined by the mass center of the two flanking base pairs, p2 and p3 are defined by the flanking sugar groups, and p4 is defined by the five-member ring of the flipping purine (or the entire six-membered ring for a flipping pyrimidine). (**C**) CPDb: a similar definition to that of CDPa, but using the phosphate groups for p2 and p3. The dotted lines show the two planes which define the pseudodihedral angles.

3.2.5 Umbrella sampling and potential of mean force calculations

The procedure for umbrella sampling was adapted from previous studies.[43] Starting from the standard B-form conformation, the initial structure of each window was generated by a 0.5 ps simulation with a restraint force constant of 10,000 kcal/(mol × radian^2) in a serial fashion, which used the previous window's last structure as the starting structure of the current window. Each window was separated by 5° from the flanking windows. After the initial structures were generated, 500 ps simulations using the same definition of restraint and a 1,000 kcal/(mol × radian^2) restraint constant were carried out for the sampling. The eversion angle and energy data were recorded at each time step. The other parameters of these simulations were the same as those of the standard MD simulations. The resulting PMF was obtained by a WHAM analysis[94-96] of the data using a program provided by Alan Grossfield (freely available at http://membrane.urmc.rochester.edu/content/wham).

3.3 *Results and Discussion*

3.3.1 Simulation of spontaneous base pair formation

Umbrella sampling simulations provide the PMF along the chosen reaction coordinates for a system. To generate an accurate free energy profile along a conformational change of interest, the reaction coordinate should be able to represent conformational change faithfully. To evaluate the extent to which various reaction coordinates can represent the process of a base-flipping event, we generated an unrestrained trajectory for a duplex in explicit water where a base pair spontaneously formed from the everted position, since it is a faster event than base opening. The starting structure was obtained from an umbrella sampling simulation using the CPD definition at −140°. A 10-ns unrestrained MD simulation was carried out. From the original everted position, the guanine spontaneously returned to the duplex and reformed the Watson−Crick base pair with the cytosine in less than 5 ns. Several snapshots taken from the trajectory are shown in Figure 3-2. In the starting structure (0 ps), the everted base (G24) is completely outside of the duplex. At ~1 ns, it moved closer to the major groove, and the purine ring adopted a conformation nearly perpendicular to the other, stacked base groups. This conformation was stable until ~3 ns, during which the purine ring attempted to reinsert into the duplex but was unsuccessful since the G24 was still in the syn conformation. It returned back to its previous everted position at ~4 ns, adopted an anti conformation at ~4.2 ns, and then successfully reinserted into the duplex at ~4.3 ns. The newly formed base pair was stable for the remainder of the 10 ns simulation (only 5 ns is shown in Figure 3-2).

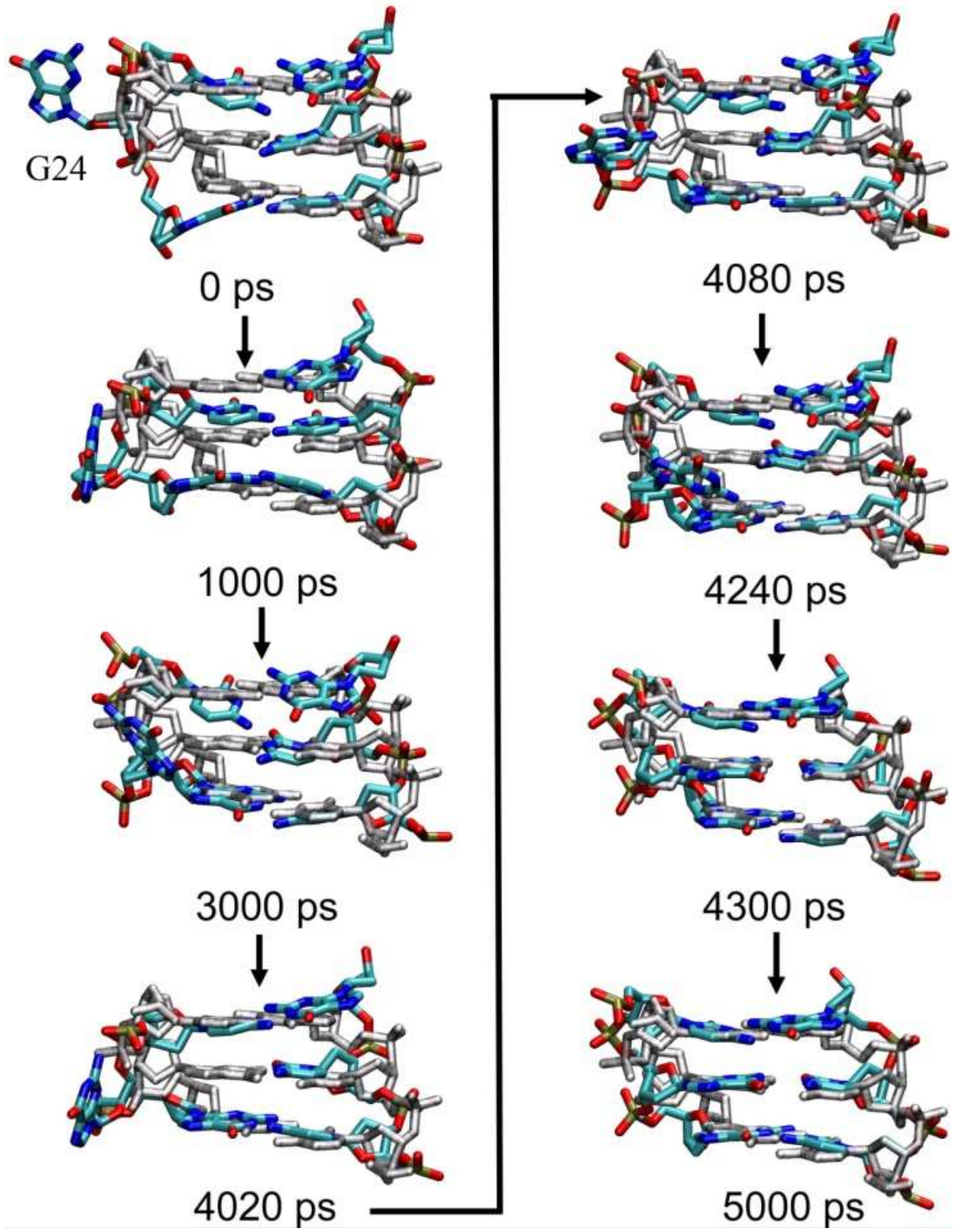


Figure 3-2. Snapshots taken from the base pair reforming trajectory viewed from the major groove. For clarity, water, hydrogen atoms, and DNA outside the central three base pairs are not shown. The simulated structures are colored by atom type. The structure in gray indicates the same duplex in a standard B-form conformation for reference. The time sequence is described in the text.

The observation of spontaneous base pair formation provides an excellent data set for evaluation of the various parameters in the base eversion restraints that will be used for umbrella sampling. In particular, any measure of base flipping should reproduce the observation that the first attempt at reinsertion by the base was unsuccessful, after which it moved back out of the major groove and then successfully inserted. We calculated several properties of the DNA duplex along this trajectory, such as the distance between the forming base pair (represented by the distance between N1 of G24 and N3 of C10), the glycosidic angle of the flipped G24 base, and the rmsd value of the central three base pairs relative to the standard B-form DNA. The results are shown in Figure 3-3.

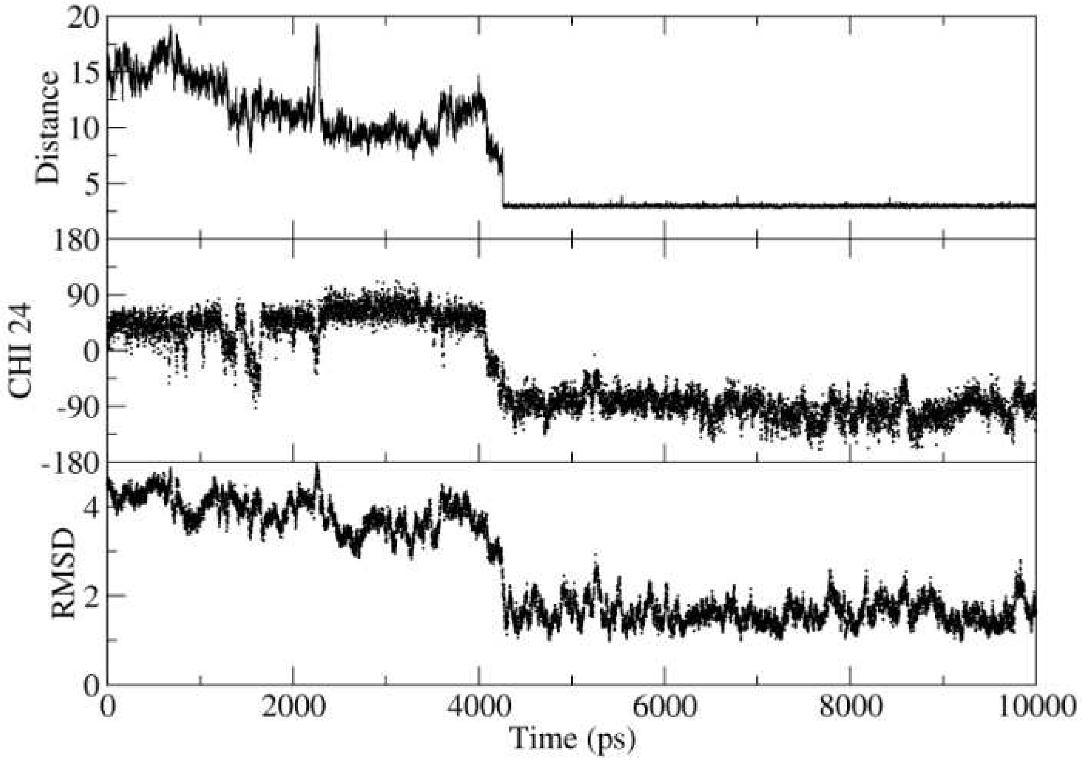


Figure 3-3. Data from the unrestrained trajectory with spontaneous base pair formation. The upper panel shows the heavy-atom-to-heavy-atom hydrogen-bonding distance between atom N1 of residue G24 and atom N3 of residue C10 between the bases in the new base pair. The middle panel shows the glycosidic angle of the flipped G24 nucleotide. The lower panel shows the rmsd of all atoms in the central three base pairs as compared to standard B-form DNA.

At the beginning of the simulation, the distance between the central two bases was ~15 Å. The flipped base G24 was in a syn conformation, with the glycosidic angle at ~55°. The rmsd value of the central three base pairs was ~5 Å compared to standard B-form DNA. The distance decreased to about 3 Å at ~4.3 ns ps; further analysis confirmed that this was accompanied by the formation of all three Watson−Crick hydrogen bonds. The reformed hydrogen bonds were stable for the rest of the simulation. Also at 4.3 ns, the glycosidic angle changed to ~ −90°, falling within the anti range, and the rmsd value decreased to ~1.5 Å, indicating that the structure was highly similar to canonical B-form DNA.

3.3.2 Evaluation of alternate base flipping reaction coordinates

The unrestrained MD trajectory exhibited a pathway of spontaneous base pair reforming, which provides an excellent data set for evaluation of the CPD reaction coordinate using COM groups as defined by MacKerell et al. and comparison to the modified approach with different selections for the COM groups (Figure 3-1). CPD denotes the base opening angle calculated using MacKerell's et al.'s center of mass pseudodihedral angle (Figure 3-1A). CPDa and CPDb are the new center of mass pseudodihedral angles (Figure 3-1B,C). During the analysis, we have found that, in the original CPD definition, the center of mass of the flipped base and the next two centers of mass can become collinear (Figure 3-4, structure images shown in Figure 3-8 and Figure 3-9). Therefore, we have also measured angles defined by these three neighboring COM positions for each CPD definition using the angle defined by points 2−3−4 shown in Figure 3-1A. Instability in the dihedral calculation should be expected if this angle approaches 0° or 180°.

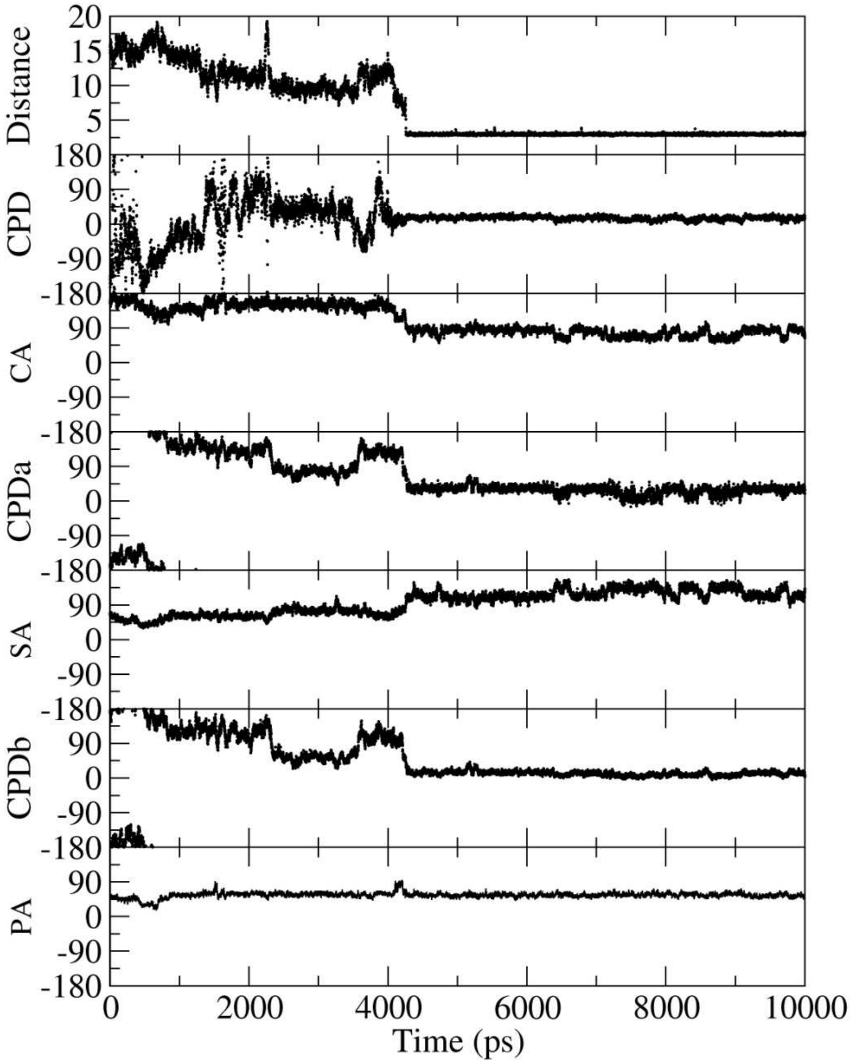


Figure 3-4. The evaluation of the reaction coordinates defined by MacKerell et al. and the new definitions. The first panel is the distance between the two bases in the base pair being formed, which is the heavy-atom-to-heavy-atom hydrogen-bonding distance between residue G24 atom N1 and residue C10 atom N3. The pseudodihedral angles are defined as the angles between two planes (see Figure 3-1). CPD is the original dihedral reaction coordinate. CPDa is the pseudodihedral angle using sugar groups. CPDb is the angle using phosphate groups. Angles between points 2, 3, and 4 for CPD, CPDa, and CPDb are shown as CA, SA (sugar angle), and

PA (phosphate angle), respectively. Instability in the dihedral results when these angles adopt values very near 0° or 180°.

From Figure 3-4, we can see that there are several potential disadvantages using the COM groups as originally defined by the original CPD angle. The first disadvantage is that the reaction coordinate does not have a one-to-one correlation between the measured and actual extent of eversion. For example, CPD-dihedral angle values sampled for everted conformations (between 2,100 and 3,100 ps) are comparable to those sampled after the base pair has formed at 4,300 ps. This means two different points on the base everting pathway will have the same value of the reaction coordinate. This is caused by the definition of the CPD angle; the last three points in the CPD definition (P2, P3, and P4, see Figure 3-1A) can become collinear in everted conformations, resulting in numerical instability. According to the original CPD definition, negative values denote flipping along a major groove pathway, and positive values indicate the minor groove pathway. While the sign of a particular flipping direction (major/minor) is arbitrary and depends on whether the CPD is defined from the 3′ or 5′ side of the flipping base, the data should be consistent once a definition of CPD is chosen. However, although the trajectory was visually confirmed to sample only the major groove pathway, the CPD during the simulation adopted both negative and positive values, suggesting that the sign is not a reliable indicator of the flipping directions.

By using the new definitions (CPDa and CPDb in Figure 3-4), the reaction coordinate values and the position of the base group have an improved correlation. The CPDa and CPDb angles gradually reduced from ~180° (extrahelical) at the beginning of the simulation to 50° at 2,300 ps. The flipped base was close to its intrahelical position at 2,300 ps, except that the base was in its syn conformation, not in the anti conformation required for proper Watson−Crick pairing (see the chi24 in Figure 3-3). Steric hindrance with the phosphodiester backbone prevents rotation about the glycosidic bond in this position; thus, the base once again moved out of the major groove, with the CPDa and CPDb correctly reflecting this change, with values increasing between 3,600 and 4,200 ps. The extra-helical base then rotated to an anti conformation and subsequently reinserted, restoring the Watson−Crick pair. This is represented with the low (~0) and steady values of the CPDa and CPDb flipping angles in Figure 3-4; we note that with the CPDa and CPDb definitions the intrahelical values (near 0) were not seen for any of the everted conformations. This is in contrast to the "intrahelical" CPD values observed at multiple points prior to the actual reinsertion event.

The second disadvantage of the original pseudodihedral definition is that one of the two angles (CA in Figure 3-4) connecting the four centers of mass can adopt values close to 0° or 180° when the base is extrahelical. From Figure 3-4, we can see that angle CA is very close to 180° before the base pair is reformed at about 4,000 ps (the distance in panel 1 of Figure 3-4 can be used as an indicator of the base pair reforming). After the base pair formed (the distance becomes a steady line at about 3 Å), the CA angle adopted values near 90°. The dihedral angles are defined by four points, where each set of three consecutive points defines a plane. The dihedral angle is the angle between these two planes. When the last three consecutive points are close to being linear (0° or 180°), a slight change of the position of the fourth point can greatly change the definition of the second plane, which results in large fluctuations in the dihedral angle (the second panel in Figure 3-4) and resulting forces. This caused unpredictable instabilities in simulations with everted bases (data not shown); CPD dihedral angle differences of ~40° were observed for nearly identical structures sampled during short time spans (Figure 3-8). In the new definitions, both angles (PA and SA in Figure 3-4) had values well away from 0° and 180° during the entire profile, resulting in improved numerical stability of the dihedral angle.

The third disadvantage of the old definition is less explicit. The definition of CPD is not symmetric (Figure 3-1), and the choice of P1 and P2 dihedral points as being either the 3′ or 5′ side of the flipping base is arbitrary. Due to the asymmetric structure of the DNA duplex, a free energy profile calculated using the two points from the 3′ side differs from that obtained with restraints for the dihedral points defined on the 5′ side. By using the new definition, the 5′ and 3′ sides are both included in a single calculation, with the resulting PMF being less ambiguous.

3.3.3 The free energy profiles calculated using the new definitions

The free energy profiles for base eversion using our two new CPDa/b dihedral angle definitions in umbrella sampling have been calculated (Figure 3-5). The upper panel shows the PMF profile using the centers of mass of the deoxyribose rings groups as dihedral points P2 and P3 (CPDa, Figure 3-1B). The lower panel shows the results using the phosphates as P2 and P3 (CPDb, Figure 3-1C). To estimate the convergence of the calculation, we also calculated the free energy profile using the second half of the data and generated error bars using the difference between the two results.

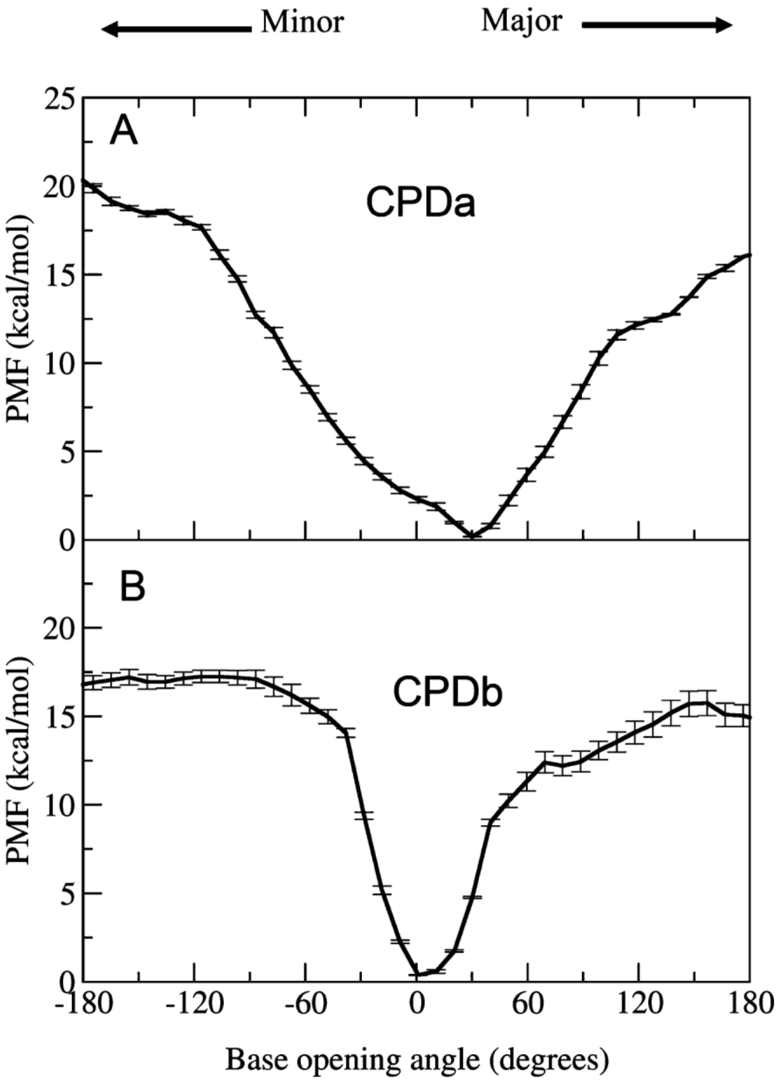


Figure 3-5. Free energy profiles using the two new pseudodihedral eversion definitions. The *x* axis is the pseudodihedral angle. Panel A shows the result using the center of mass of the sugar rings as points P2 and P3 (CPDa). Panel B shows the result using the center of mass of the phosphates as dihedral points P2 and P3 (CPDb). Positive/negative values reflect flipping into major/minor grooves, respectively. The solid line shows the free energy profile calculated using the last 400 ps of a total 500 ps per window. The error bar shows the difference of the results calculated using the last 400 ps data and the last 200 ps data.

From Figure 3-5, we can see that there are certain similarities and differences between the two energy profiles. Both profiles can be divided into two regions: the “basin” and the “plateau” regions. The basin region is near the energy minimum and has a lower free energy and steeper slope. The plateau region reflects everted bases and is further from the minimum with a high free energy value and less energetic dependence on the angle. The basin region for the profile using the phosphate groups (CPDb) was between −45° and +45°. The basin using the sugar group for the PMF reaction coordinate (CPDa) is significantly broader, ranging from −120° to +120°. In both cases, the barriers of the PMFs (after the basin region ends) at the major groove pathway were lower than those at the minor groove pathway. The values for using the sugar rings were about 13 kcal/mol for the major groove and 18 kcal/mol for the minor groove. The values for using phosphate groups were about 12 kcal/mol for the major groove and 15 kcal/mol for the minor groove. These results agree reasonably with calculations by Banavali and MacKerell, showing that the energy barriers for G flipping are 18.7 and 21.3 kcal/mol for the major and minor groove, respectively.[43] The positions of the minima and height of the energy barriers of Benavali and MacKerell’s and our studies are similar but do not exactly match. One possible reason is that the sequence contexts are different between these two studies. However, both experimental and theoretical studies have suggested that base opening rates have little dependence on sequence context.[140,146] The differences may also be due to the influence of the reaction coordinate definition on the PMF details, or the enforcement of periodicity in the free energy calculation, or the difference between CHARMM and Amber force fields. The present results are also comparable to the Amber results of Priyakumar and MacKerell,[49] though that study used an older version of the Amber DNA force field than used here.

Although the free energy barriers were similar in the two free energy profiles in Figure 3-5, the widths of the basin regions were significantly different. The basin is much wider for the CPDa definition. To understand why the free energy profiles are different, we calculated the correlations between CPDa and CPDb definitions for structures sampled in the two umbrella sampling runs, which are shown in Figure 3-6. Data from both simulations have a high correlation between CPDa and CPDb values. However, both correlation figures are not straight lines. There is a flat phase near 0°, which corresponds to the intrahelical conformation of the base. In other words, the intrahelical space is wider using CPDa as reaction coordinates than using CPDb, even for the same structure sets. This is due to the difference in the geometry of the two definitions in the intrahelical region. In Figure 3-4, we can see that CPDa is has obtuse angle SA, and the PA of CPDb is ~60° when the base is in its intrahelical conformation. Since it corresponds to the intrahelical conformation, the basin region is wider in CPDa space.

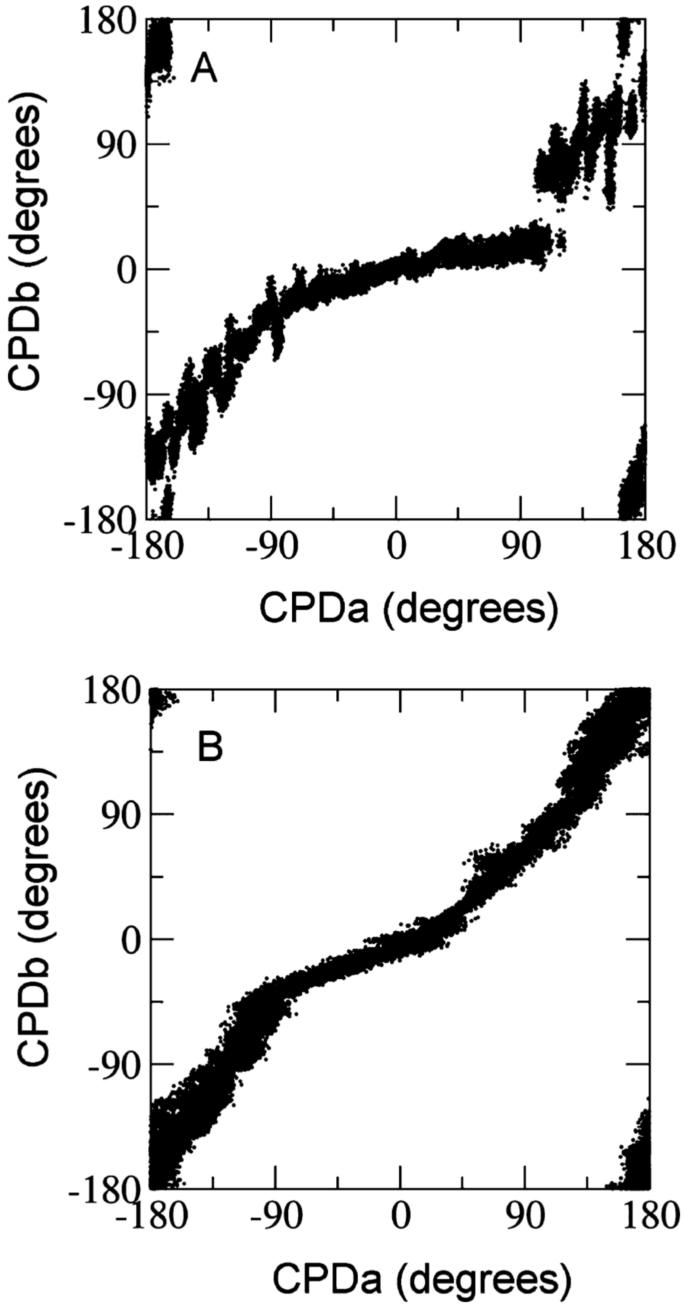


Figure 3-6. The correlation of the two CPDa/b definitions. Panel A shows the structures sampled using the CPDa restraint. Panel B shows the structures sampled using the CPDb restraint. The *x* axis shows the postprocessing results using CPDa, while data for CPDb are on the *y* axis. Regardless of the restraint used to generate the structures, CPDa is more sensitive than CPDb to changes in the region near 0°.

We further analyzed the correlation between flipping angle and base pair distance to investigate the properties of these two reaction coordinate definitions (Figure 3-7). Figure 3-7A shows the structures sampled in the umbrella sampling simulation using the CPDa restraints. Figure 3-7B shows the structure sampled in the umbrella sampling simulation using CPDb restraints. For all structures in each simulation, we calculated the distance between the flipping base and its partner, as well as flipping angles measured using both dihedral definitions. As we observed with the comparison of the two dihedral angles in Figure 3-6, data from simulations performed with either definition as the restraint are consistent. In both cases, the base pair distance was about 3 Å and stable for the region around a base opening angle of 0°. In both simulations, the region of close contact between the bases covers a significantly larger range when using CPDa as a reaction coordinate than when using CPDb as a reaction coordinate, even when they were applied to the same set of structures. This confirms that CPDb is more sensitive to the true extent of base opening. We can also see that the structures sampled using CPDb (Figure 3-7B) are more similar among windows than the ones sampled using CPDa (Figure 3-7A). The structures sampled using CPDa (Figure 3-7A) seem more poorly converged. The reason CPDa and CPDb behave differently may be due to the number of internal coordinates encompassed in the two definitions. CPDb includes two phosphate groups and one sugar ring between P2 and P3, while CPDa includes two phosphate groups and three sugar rings (see Figure 3-1). Since the CPDa definition has a more complex conformation space, it is more sensitive to structural fluctuations in the backbone.

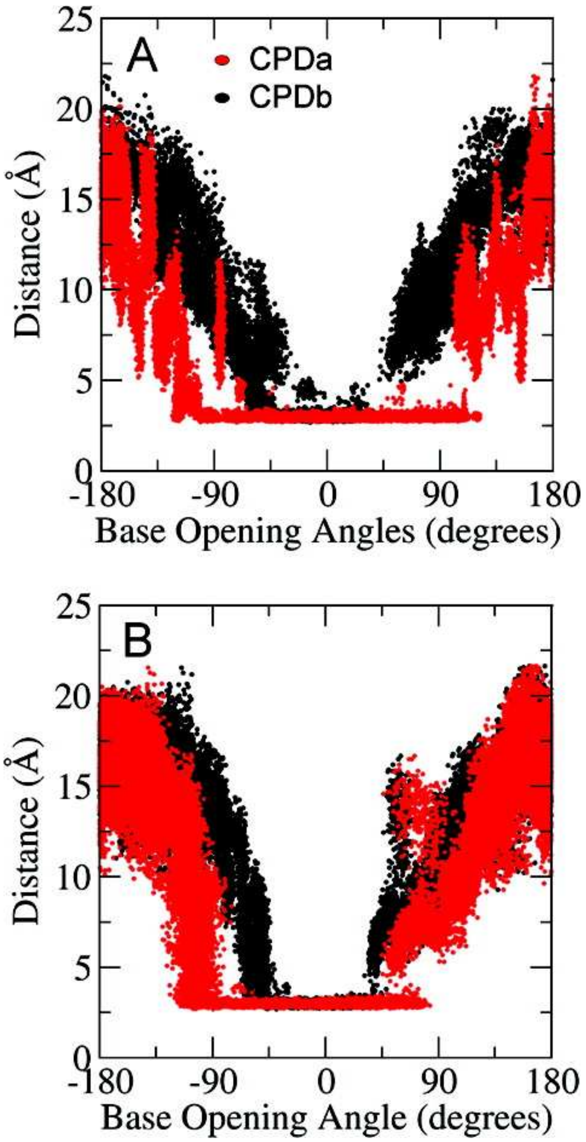


Figure 3-7. The correlation between the distance of the two bases (distance between atom N1 of residue G24 and atom N3 of residue C10) and the CPDa/b base opening angles. The upper figure shows the structures sampled in the umbrella sampling using CPDa as the reaction coordinate. The lower figure shows the structures sampled in the umbrella sampling using CPDb as the reaction coordinate. For each simulation, the data shown in red are the postprocessing results using the CPDa definition, and those in black are the results using CPDb. Both plots indicate that the CPDa definition has a broader range of values for base-paired structures as compared to CPDb, independent of which restraint was used to generate the flipping trajectory.

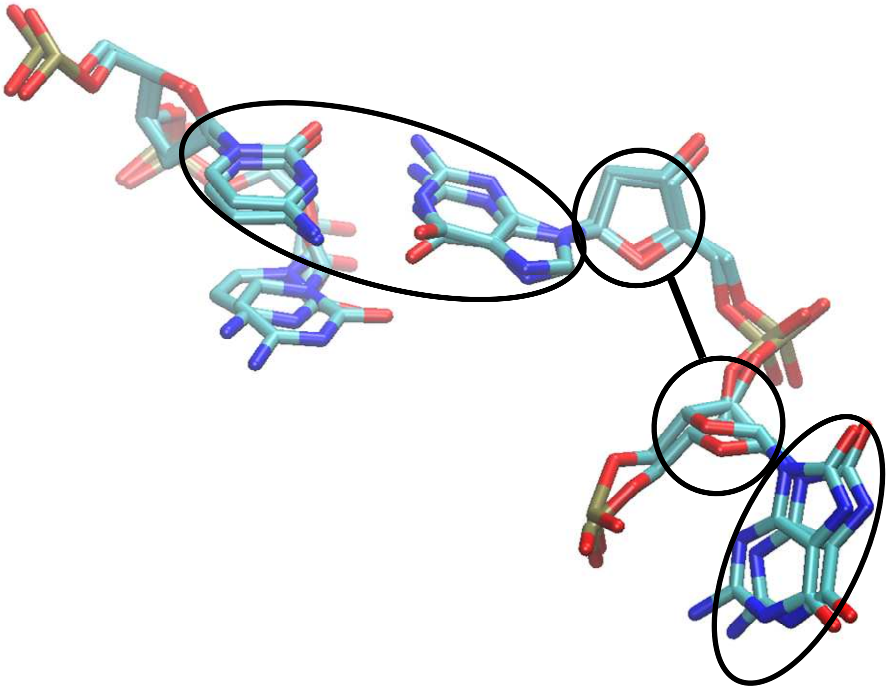


Figure 3-8. Overlap of two highly similar structures sampled using a CPD dihedral angle restraint at the 170° window. The four regions which define the CPD dihedral angle are circled; dihedral values for the 2 structures shown are 161° and 78° due to the nearly collinear angle for the rightmost 3 groups. This results in simulation instability.

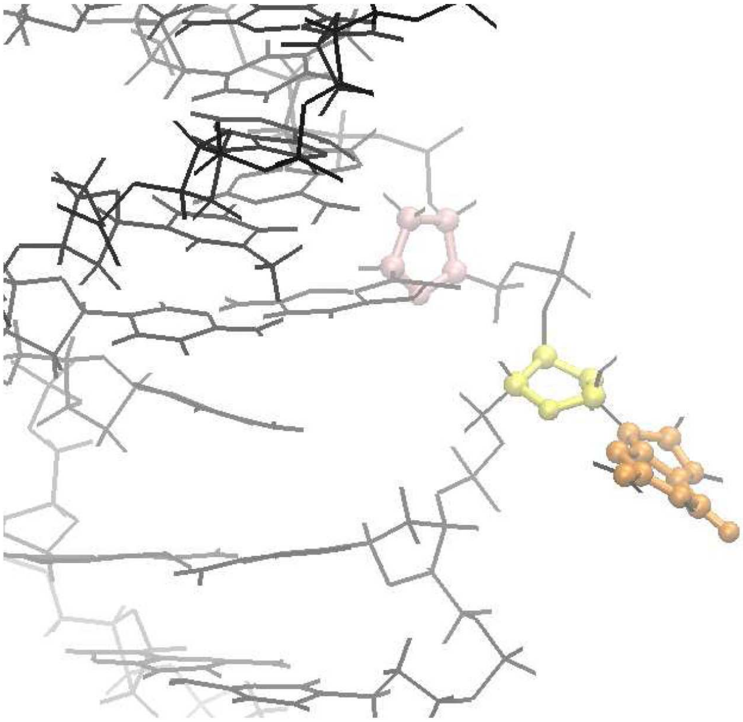


Figure 3-9. Additional view of an everted DNA structure in unrestrained simulation with 3 nearly collinear groups in the CPD dihedral definition (atoms in COM groups are shown as colored spheres).

3.4 *Conclusions*

Base flipping is an important event, and computational tools have been shown to be essential in studying processes such as an enzymatic role in flipping. Starting from a flipped conformation, we have generated a fully unrestrained MD trajectory in which an everted guanine base spontaneously returned to its intrahelical conformation and reformed its Watson−Crick pair with the cytosine partner. This trajectory was used to evaluate a previously proposed pseudodihedral angle and how well it describes the extent of eversion. We found several disadvantages in the definition, including the potential for numerical instability, and used the data to propose two modified pseudodihedral definitions which can successfully avoid the observed disadvantages. The free energy profiles of the base flipping using the new definitions have been calculated, and the results reasonably agree with previously published results. We also compared the two modified definitions. The one using the center of mass of the phosphate groups has a tighter correlation with the base opening angle; therefore, it is a better representation for base flipping. The reaction coordinate using the center of mass of the sugar groups has a larger conformation space and appeared to be more difficult to use in generating well-converged data. In closing, we remind the reader that using any restraint to impose a reaction coordinate may introduce artifacts in the data as compared to fully unrestrained systems.

Chapter 4 Active destabilization of damaged base pairs by a DNA glycosylase wedge

Abstract

The BER glycosylases share structural traits when recognizing DNA damage. One particular functionality that BER glycosylases share is using aromatic or aliphatic residue to probe the DNA and destabilize the base pair of interest to facilitate eversion.[100] This behavior is seen with the bacterial glycosylase Fpg, in which a phenylalanine (113) wedge is used to probe for damage. In this work, we investigate the role that this Phe113 wedge plays in destabilizing the 8-oxo-G:C base pair to facilitate eversion. To accomplish this, computationally we mutated the Phe113 wedge to an alanine to better understand how the removing the intercalating wedge affects the overall profile to eversion. To measure the energetic profiles to base eversion for 8-oxo-G and the 8-oxo-G.F113A, partial nudged elastic band method was coupled with umbrella sampling.[91] After obtaining the energetic profiles to eversion, we found that removing the wedge stabilizes the intrahelical state. Through the use of structural analysis and pairwise energy decomposition, we uncover the key interactions that result from the the presence of the intercalating wedge.

Acknowledgments

The material presented in this chapter contains direct excerpts from a manuscript that is in preparation by Nikita A. Kuznetsov, Arthur J. Campbell, Christina Bergonzo, Grigory V. Mechetin, Vladimir V. Koval, Carlos de los Santos, Arthur P. Grollman, Dmitry O. Zharkov, Olga S. Fedorova, and Carlos Simmerling. The main text of the manuscript was written by Professor's Dimitry Zharkov and Carlos Simmerling. The methods section was written in part by Arthur J. Campbell with suggestions and revisions from Professor Carlos Simmerling. Nikita A. Kuznetsov, Grigory V. Mechetin, Vladimir V. Koval, Dmitry O. Zharkov and Olga S. Fedorova conducted the experimental studies. Arthur J. Campbell, Christina Bergonzo and Lin Fu performed the simulations and data analysis.

4.1 Introduction

DNA repair enzymes must efficiently discriminate between normal and damaged DNA elements, which are often very similar chemically. Formamidopyrimidine-DNA glycosylase (Fpg) is a repair enzyme that excises oxidatively damaged purines from DNA. Its main substrate bases include two derivatives of guanine, 8-oxo-Guanine (8-oxo-G) and 2,4-diamino-6-oxo-5-formamidopyrimidine (Fapy-G), as well as the formamidopyrimidine derivative of adenine. The structural basis of DNA binding by Fpg enzymes from several bacterial species, including *Escherichia coli*, *Lactococcus lactis* and *Bacillus stearothermophilus*, has been revealed by X-ray crystallography [66,70,147-150] and molecular dynamics (MD) [151-155]. As with all other DNA glycosylases investigated to date, the complexity of substrate structure requires extensive conformational changes in DNA after binding to the enzyme to achieve a catalytically competent conformation. These changes include kinking of the DNA duplex and eversion of the damaged deoxynucleoside from the body of the duplex into the lesion-binding pocket of the enzyme. The enzyme molecule, in turn, also undergoes several conformational transitions, among them possible closing movement of the enzyme domains, insertion of several amino acid residues into the void vacated by the everted damaged deoxynucleoside, and isomerization of the active site pocket. The time course of Fpg-catalyzed excision of 8-oxo-G, studied by stopped-flow methods[156-158], reveals at least five conformational transitions in the enzyme-substrate complex likely preceding the chemical steps of the reaction. As a result of this multistage mutual conformational adjustment of the enzyme and its substrate, Fpg discriminates very strongly in favor of 8-oxo-G and against undamaged G.

In a complex with damaged DNA after base excision, both *Bst*-Fpg[138] and *Eco*-Fpg[66] wedge a Phe side chain (Phe110 in *Eco*-Fpg, Phe113 in *Bst*-Fpg; Phe110/113) between a damaged base pair and a pair 5′ to the damaged base. In addition, two side chains (Met73/76 and Arg108/111) are inserted into the void vacated by the everted damaged dN, and the axis of DNA is severely kinked (66–75°). Two groups of the *Bst*-Fpg structures represent pre-excision stages, one of Fpg covalently cross-linked to undamaged DNA, another of Fpg (an E2Q catalytically inactive mutant) in complex with 8-oxo-G-containing DNA.[71,159] In a complex with normal DNA, Phe114 of *Bst*-Fpg is wedged into the DNA duplex next to the undamaged G:C pair, which is notably buckled, and the DNA is kinked, but no eversion of the normal base is observed. In contrast to these gross changes in DNA (and possibly protein) structure, a stopped-flow analysis of binding of *Eco*-Fpg to undamaged DNA surprisingly reveals only a single binding stage in the time course traces produced by fluorescence of the enzyme’s native Trp residues or of 2-aminopurine (2-aPu) reporter bases incorporated into the substrate[156,157,160]. In principle, detection of conformational transitions at early stages of DNA binding could be hindered by lack of fluorescent groups appropriately placed at the enzyme–DNA interface. On the other hand, it is not clear whether the “insertion-before-eversion” structure[71] represents a true Fpg intermediate or an experimental artifact due to a covalent stabilization of the enzyme–DNA complex.

High-quality atomistic simulation, based on available three-dimensional structures, is being increasingly used to theoretically analyze conformational changes in enzymes during their catalytic cycle.[161] On the other hand, transient enzyme kinetics supplemented with site-directed mutagenesis and substrate perturbation provides valuable experimental information regarding a potential role for dynamics, but usually does not allow atomic resolution.[162] Combining atomistic simulation with kinetics experiments may provide an insight into the dynamic nature of substrate recognition and processing by enzymes. We have used such a combined approach to address the mechanism of a critical step in substrate recognition by Fpg, initial destabilization of the duplex at the interrogation site. Crystallographic structures of Fpg have been reported for *B. stearothermophilus* (*Bst*-Fpg), and were used as starting points for MD simulations. The complementary transient (stopped-flow) kinetic experiments were performed with the *E. coli* enzyme (*Eco*-Fpg), which is better characterized biochemically, and for which stopped-flow data are already available.[156,157,160]

4.2 Methods

4.2.1 Initial structure generation for 8-oxo-G.wt and 8-oxo-G.F113A

The Amber 9 and 10 suites of programs were used for all molecular mechanics calculations in this work.[120] Several systems were simulated: wild type Fpg bound to DNA interrogating 8-oxo-G (8-oxo-G.wt), Fpg with a F113A mutation bound to DNA interrogating 8-oxo-G (8-oxo-G.F113A), Fpg with a F113W mutation bound to DNA interrogating 8-oxo-G (8-

oxo-G.F113W) and unbound duplex DNA containing an abasic site (DNA.AP). For all systems in this work the ff99SB force field was used with the parmbsc0 DNA backbone parameters.[81,119]

Initial structures for the intrahelical and extrahelical states were generated from two crystal structures, pdb IDs 2F5O and 1R2Y (Fpg interrogating an intrahelical and extrahelical base respectively).[70,71] In the 2F5O crystal structure, Fpg is bound to DNA and is interrogating an intrahelical normal Watson Crick G:C base pair.[71] In the 1R2Y crystal structure, Fpg is bound to DNA which contains an 8-oxo-G that is fully extrahelical and in the active site.[70] Since there are differences between the crystal structures (described further below) both structures were modified to be equivalent in their primary structure. The DNA sequence of 2F5O was used as the standard (Table 4-3). Any disordered base in the 2F5O sequence was built in using MoilView from a standard duplex made in nucgen. Any base that was not identical in 1R2Y to the sequence in Table 4-3 was deleted, keeping the sugar and DNA phosphate. The correct atoms for Table 4-3 were then built using Amber's tleap module.[27] Missing nucleotides were modeled into the 1R2Y structure using MoilView. The 2F5O crystal structure does not have an ordered active site loop. This was modeled using the ordered active site loop from the 1R2Y crystal structure. The E3Q inactivating mutation in 1R2Y was reversed to the wildtype E3 sequence. Although the experimental conditions included protein-DNA crosslinks, these were not included in the published coordinates. The simulations corresponded to the un-crosslinked species. Conversion of G to 8-oxo-G was performed using Amber's tleap module.

Table 4-3. DNA sequence and residue numbering of the 8-oxo-G.wt, G.wt, 8-oxo-G.F113A, 8-oxo-G.F113W systems used in this work.

274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	
A 5'	G	G	T	A	G	A	T	C	C	G	G	A	C	G	C 3'	
	C 3'	C	A	T	C	T	A	G	8OG	C	C	T	G	C	G	T 5'
	305	304	303	302	301	300	299	298	297	296	295	294	293	292	291	290

4.2.2 Structure equilibration

After both intra- and extra-helical structures were mutated to chemically equivalent systems, they were minimized, equilibrated and relaxed further in unrestrained dynamics.

Two sequential rounds of minimization were run for 200 steps of steepest descent where all heavy atoms of the systems were held in place with a restraint force constant of 5.0 kcal/mol-Å^2. After minimization the systems were then solvated with ~10,000 TIP3P explicit solvent water molecules in a truncated octahedron periodic box with an 8Å buffer. After the systems were solvated they were further minimized over five consecutive 1,000 step rounds with the steepest descent method where the restraint force constant was incrementally relaxed from 10.0, 5.0, 2.5, 1.0 and 0.0 kcal/mol-Å^2 respectively on all atoms of the DNA and Fpg.

Five consecutive rounds of equilibration were employed after minimization under constant pressure at 1 atm with a timestep of 2fs where all bonds involving hydrogen were constrained using SHAKE. Temperature was controlled by using the weak-coupling algorithm. In the first round of equilibration all atoms of the DNA and enzyme were frozen while all water molecules were allowed to relax for 50ps. Also during the first step of equilibration the system was slowly heated from 0 to 300K. In the second, third and fourth steps of equilibration the temperature was held constant at 300K. In the second step of equilibration all atoms on the DNA and Fpg were restrained with a restraint force of 5.0 kcal/mol-Å^2 for 50ps. In the third step the restraints on the DNA and Fpg were decreased to 1.0 kcal/mol-Å^2 for 50ps and in the fourth step the restraints were completely removed and the entire system was relaxed in unrestrained dynamics for 50ps. In the fifth and last step the system was slowly heated from 300K to 330K over the course of 500ps since Fpg is a Thermaphile and 330K is the biologically relevant temperature. After the last step of equilibration the system was simulated for ~20ns in unrestrained dynamics. The system was coupled to a thermostat and barostat at 330K and 1atm, respectively. After the intrahelical and extrahelical structures were equilibrated the final structures were used as starting structures for the remainder of the simulations.

4.2.3 Simulation details of abasic site

A B-form DNA was built in Nucgen and mutated a central base into an abasic site by deleting the base and using tleap to build in any missing atoms.[27] Next the system was solvated with ~8,000 explicit solvent TIP3P water molecules in a truncated octahedron periodic box with an 8Å buffer. Two consecutive rounds of minimization were run. The first round of minimization was run for 500 steps of steepest descent followed by 500 steps of conjugant gradient minimization while restraining all atoms in the DNA with a restraint force constant of 500.0 kcal/mol-Å^2. In the second round of minimization all restraints were removed and the minimization was run for 1,000 steps of steepest descent followed by 1,500 steps of conjugant gradient. The minimized structure was then run in through two round of equilibration. In the first round of equilibration the system was then heated to 330K over the course of 20ps while all atoms on the DNA were restrained with a restraint force constant of 10.0 kcal/mol-Å^{2}a temperature of 330K was chosen to increase the sampling of the system. All atoms involving hydrogen were constrained using SHAKE and a time step of 2fs was used for the remainder of these simulations. In the section round of equilibration, all restraints were removed and the system was able to relax for 100ps. After equilibration the system was simulated for 5ns at 330K and1atm. The last structure of this simulation was used in Figure 4-10 analysis.

4.2.4 Running PNEB

The protocol for simulated annealing used in this work was adopted from Mathews & Case 2006.[163] Six steps of heating, cooling and equilibration were used. To sample major groove eversion paths, midpoint beads were used to guide the everting base in the proper direction. The initial beads for 8-oxo-G.wt were generated from the endpoint structures discussed in the previous section, as well as midpoint structures which were partially everted to guide the system through the major groove(discussed further below).

Partially everted structures were required to seed the eversion path of 8-oxo-G.wt through the major groove. To develop the midpoint beads for 8-oxo-G.wt, a short restrained molecular dynamics simulation was run where the 8-oxo-G was forced to evert out through the major groove from the intrahelical position to an eversion angle of ~195 degrees. This simulation was run for 30ps with a timestep of 2fs. The temperature was maintained at 330K using the weak coupling algorithm. The backbone atoms of the Fpg were restrained with a .10 kcal/mol-Å^2 force constant. For the DNA, the entire strand opposite the everting base strand was restrained along with residues 293-297 (Table 4-3) with a force constant of .10 kcal/mol-Å^2. On the restrained DNA residues, all heavy atoms of the sugar and the phosphorus atom were restrained. A force constant of .0305 kcal/mol-degrees^2 was used to evert the 8-oxo-G base (Table 4-3, residue 297) out of the duplex and in the positive Phi direction. The 30 frames from this trajectory along with the intrahelical and extrahelical endpoint beads from the previous section II were used as the initial 32 beads that were initially used for PNEB.

In PNEB, the initial path must be optimized to find the minimum energy pathway. Therefore, six consecutive rounds of annealing were performed on the 32 initial beads described above. Springs were used to keep images evenly spaced along the path and a decoupling by a tangent based on neighbor bead positions or energies removed the effect of the springs on the optimization of each bead by the force field. The PNEB code in Amber's multisander functionality allowed us to apply NEB forces to a subset of the system's atoms.[27,91] For all systems discussed here, NEB forces were applied to all heavy atoms of the protein and DNA. All bonds involving hydrogen were constrained using SHAKE and volume was held constant. A Langevin thermostat and a nonbonded cutoff of 8.0 Å was used in all six steps of annealing.

In the first round of annealing for PNEB, all 32 beads were simulated for 80ps with a timestep of 2fs. The spring constant was 10 kcal/mol-Å^2. The system temperature was held at 330K with a collision frequency of 100.0 ps^{-1}.

The spring constant was then increased from 10 to 50 kcal/mol-Å^2 in the second round of PNEB annealing. Each bead was simulated for 500ps with a timestep of 1fs at 330K with a Langevin collision frequency of 75.0 ps^{-1} which was held constant for the remaining rounds of annealing.

Next the spring constant was decreased from 50 to 25 kcal/mol-Å^2 for the third round of annealing. In this stage the system was incrementally heated from 330K to 386K over the course of 60ps with a timestep of 2fs.

In the next stage of PNEB the temperature of the system was held constant at 386K for 220ps with a timestep of 1fs. The spring constant in this round was held at 25 kcal/mol-Å^2.

In the fifth round of annealing the spring constant was still kept at 25 kcal/mol-Å^2 while the system was incrementally cooled from 386K to 330K over the course of 100ps with a timestep of 2fs.

In the final round of PNEB the system was equilibrated at 330K over the course of 500ps with a timestep of 2fs and a spring constant at 25 kcal/mol-Å^2.

4.2.5 Umbrella Sampling

The Umbrella Sampling technique was employed to obtain free energy profiles of eversion. After running the final round of PNEB, which developed low potential energy paths going from the intrahelical to the extrahelical position, the trajectories were decomposed into two reaction coordinates, the eversion angle (Figure 4-6) and the glycosidic torsion angle. The starting structures for Umbrella Sampling were taken from this last step of PNEB, by extracting the closet structures based on a predefined grid spacing of 10 degrees with a limit of 10 degrees from the structure to the closet grid point (figure S8, top right) (~300 total windows) which represents each starting structure for each separate Umbrella Sampling window. Each window was restrained to its respective grid point coordinates with a restraint force constant of .0609 kcal/mol-degrees^2.

The 8-oxo-G.wt system was simulated for 500ps with a timestep of 2fs. The temperature of each window was held constant at 330K using a langevan thermostat with a Langevin collision frequency of 75.0 ps^{-1} and volume was held constant. An 8Å nonbonded cutoff was applied to all windows. All bonds involving hydrogen were constrained using SHAKE.

A force constant of .0609 kcal/mol-degrees^2 gave an optimal amount of overlap between neighboring windows, (Figure 4-7, bottom right). The eversion and Gly angles were written for each timestep.After the Umbrella Sampling simulations were complete the two dimensional

Potential of Mean Force (PMF) profiles were extracted from the eversion and glycosidic data written using 2D-WHAM analysis (Figure 4-7, bottom left) (freely available at http://membrane.urmc.rochester.edu/).

4.2.6 8-oxo-G.F113A mutation details

Since we were interested in the mutation along the identical low energy path sampled in the 8-oxo-G.wt system from PNEB, the F113A mutation was based on all ~300 initial starting structures that were extracted from PNEB as initial structures for 8OG.wt. To execute these mutations, all atoms on the Phe side chain were deleted while keeping the backbone atoms and using Tleap to change the residue to an Ala.

Each window was equilibrated for 25 ps with a timestep of 2 fs. All heavy atoms of the DNA and Fpg were restrained with a restraint force constant of 10 kcal/mol except residue 113 which was allowed to relax along with the explicit solvent. In this stage the temperature was held constant at 330K using a Langevan thermostat with a collision frequency of 75.0 ps^{-1}. After equilibration the same protocol was used that is outlined above for 8-oxo-G.wt Umbrella Sampling to obtain PMF profiles.

To calculate the Potential of Mean Force (PMF) profiles for 8-oxo-G.wt and 8OG.F113A two Dimensional Weighted Histogram Analysis (2D-WHAM) was employed. The basic raw data input for 2D-WHAM in this work was the eversion and glycosidic torsion angles at each time step for each Umbrella Sampling window along with its respective restraint force constant holding each window to its assigned grid point and the temperature the Umbrella Sampling calculations were simulated at (330K).

The PMF profiles of Figure 4-11 (top and middle) were generated using Origin 8.0 edition where the extrahelical position was set to zero to compare the relative stability differences between 8-oxo-G.wt and 8-oxo-G.F113A. To further clarify the differences between 8-oxo-G.wt and 8-oxo-G.F113A we subtracted 8-oxo-G.wt minus 8-oxo-G.F113A (Figure 2, bottom), any positive region signifies that the 8-oxo-G.F113A system is more stable.

The data for Figure 4-3 and Figure 4-4 were binned in the same manner. Each eversion angle, glycosidic torsion angle and variable data point was analyzed for each picosecond of all the umbrella sampling data. Each bin was the average variable data point for a 5 by 5 degree bin.

The initial data for Figure 4-11 and Figure 4-15 were binned separately in the same manner as above for 8-oxo-G.wt and 8-oxo-G.F113A and then subtracted (8-oxo-G.wt minus 8-oxo-G.F113A) to show the relative stabilities.

4.2.7 8-oxo-G.F113W mutation details

Several of our FRET experiments included the F113W mutation. To validate that the tryptophan is a reasonable intercalating analog, we modeled in the F113W mutation in the 8-oxo-G.wt system. After the last step of equilibration of the intrahelical 8-oxo-G bound to Fpg we mutated F113W. The backbone and part of the F113 was kept to maintain the planity in the mutation. After the mutation we minimized the structure for 1,000 steps of steepest descent followed by 20ns of unrestrained dynamics under the same conditions outlined above for 8-oxo-G.wt.

4.2.8 Energy decomposition

To measure the energy decomposition MMGBSA was employed, using the recently implemented version of sander where idecomp was set to 2 or 4. The internal energy (idecomp=2) of residue 297 (Figure 4-16-A) was measured which includes the internal (bonds, angles, dihedrals) energies as well as the 1-4 electrostatics and van der Waals interactions. The pair wise per-residue energy was measured for residue 297 to 298 and residue 297 to 283 and residue 297 to 76 (Figure 4-16 B to D), this includes the electrostatic and van deer Waals nonbonded interaction terms.

4.2.9 Parameters

The parameters used in this work for 8-oxo-G were based on the parameters developed by Miller et al.[164] The alpha and gamma backbone parameters were altered to include the parmbs0 parameters for DNA.[81]

The parameters used in this work for the abasic site were created in the Simmerling laboratory. The parameters for the abasic site was generated in two parts; the first part was the charge fitting and the second was atom type assignment.

The charges for the abasic site were generated from three alternate conformations and fit using RESP fitting method.[117] All other parameters we chosen through analogy. The atom types were assigned through analogy.

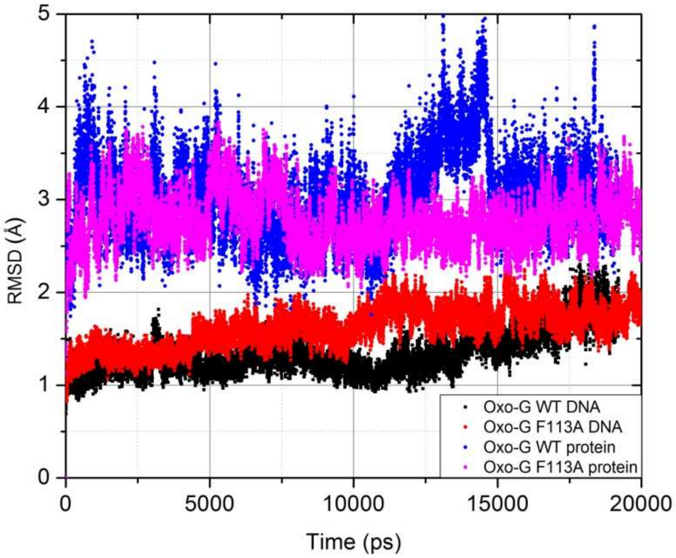


Figure 4-1. RMSD vs. time for unrestrained intrahelical 8-oxo-G.wt and 8-oxo-G.F113W with F and W wedges respectively. The reference structure for both trajectories is the first structure. Black and red represents the RMSD of 8-oxo-G WT and 8-oxo-G.F113W heavy atoms on the DNA. Blue and purple represents the RMSD of 8-oxo-G.WT and 8-oxo-G.F113W backbone atoms of the protein.

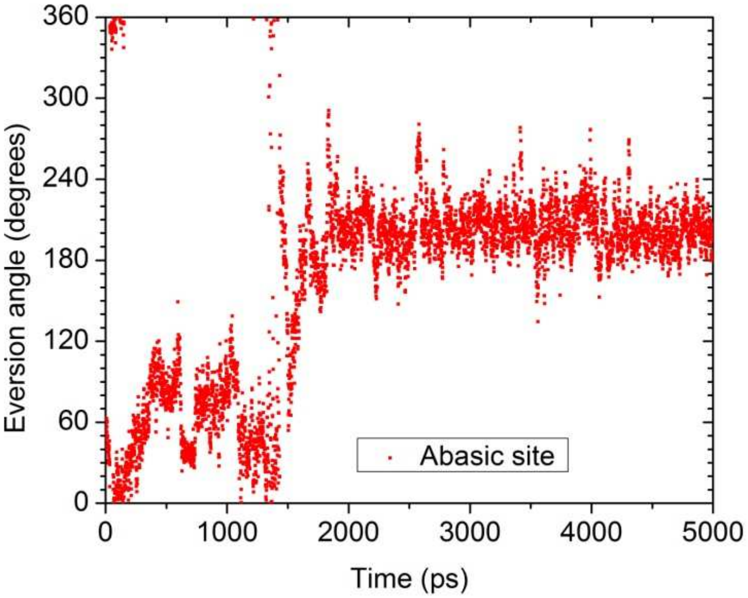


Figure 4-2. Eversion angle as a function of time for duplex DNA containing an abasic site. The initial intrahelical form undergoes a spontaneous transition through the major groove to an averted position with eversion angle phi of ~200 degrees. The eversion angle was measured in a similar manner as shown in Figure 4-6 with the exception of point 4 which includes only the O1B atom.

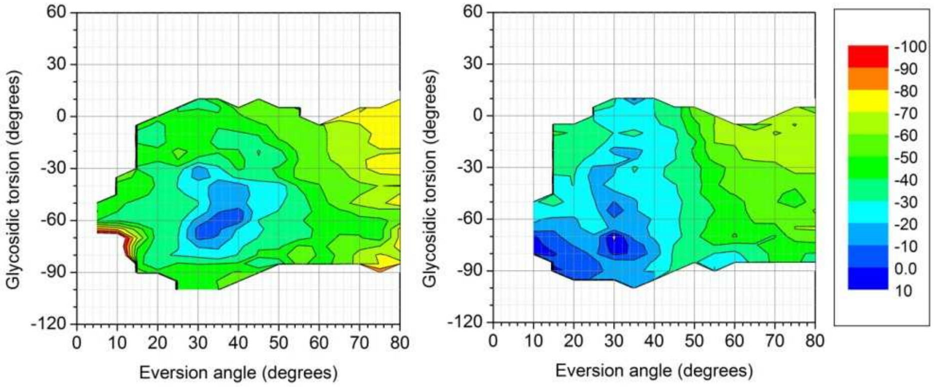


Figure 4-3. **.** 8-oxo-G:C buckling during the initial stages of eversion (eversion angle up to 80° is shown) for WT (left) and F113A (right) Fpg. Significant increases in buckling are observed in the presence of the Phe wedge in the intrahelical region which is an eversion angle less than 30 degrees.

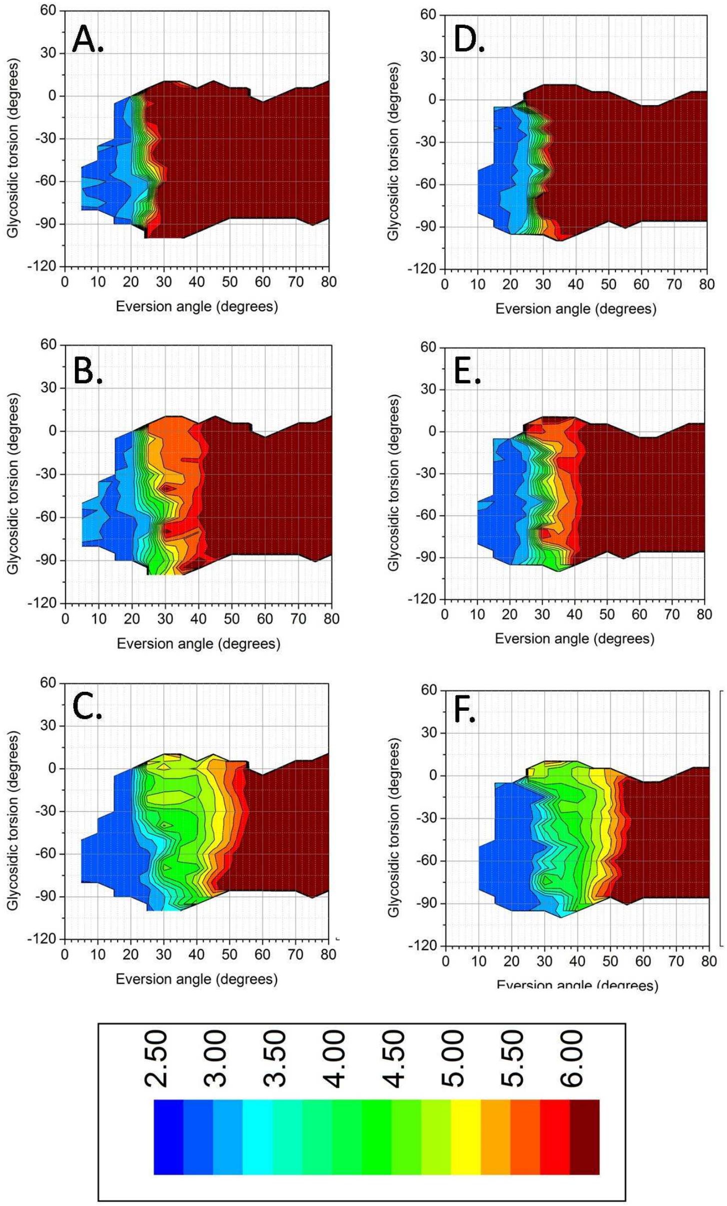


Figure 4-4. Distances corresponding to the three 8-oxo-G:C Watson Crick hydrogen bonds during eversion. The left column shows data for WT Fpg, and the right shows F110/113A wedge deletion mutant data. In each case, the hydrogens bonds are seen to break at lower extent of eversion (at lower eversion angles). Atom names correspond to Figure 4-5. A/D corresponds to the O6 to N4 distance, B/E corresponds to the N1 to N3 distance and C/F corresponds to the N2 to O2 distance between 8-oxo-G and C respectively.

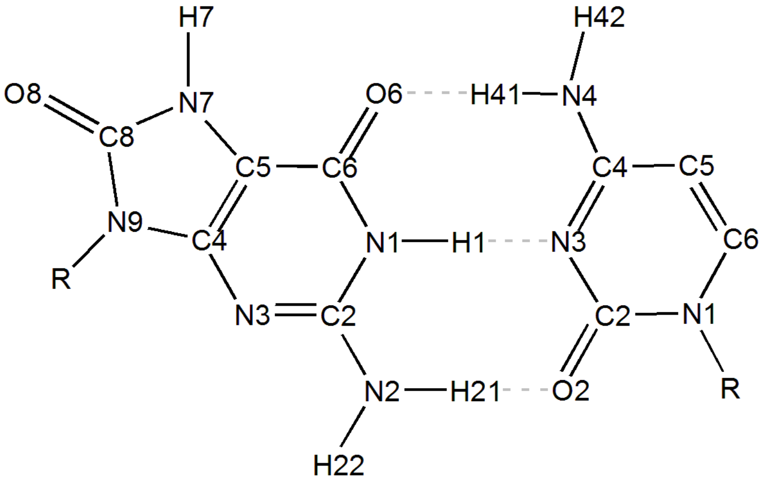


Figure 4-5. Atom names used for 8-oxo-G (left) and C (right)

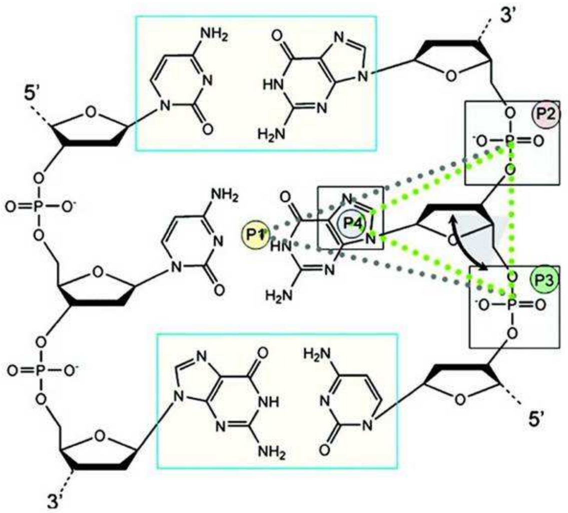


Figure 4-6. Angle of base eversion. The eversion angle is separated into four points: point 1 is the center of mass (COM) of the heavy atoms of the two flanking base pairs to the everting base, point 2 is the COM of the 3' phosphate to the everting base, point 3 is the COM of the 5' phosphate group to the everting base and point 4 is the COM of the heavy atoms of the 5 member ring of the everting base.

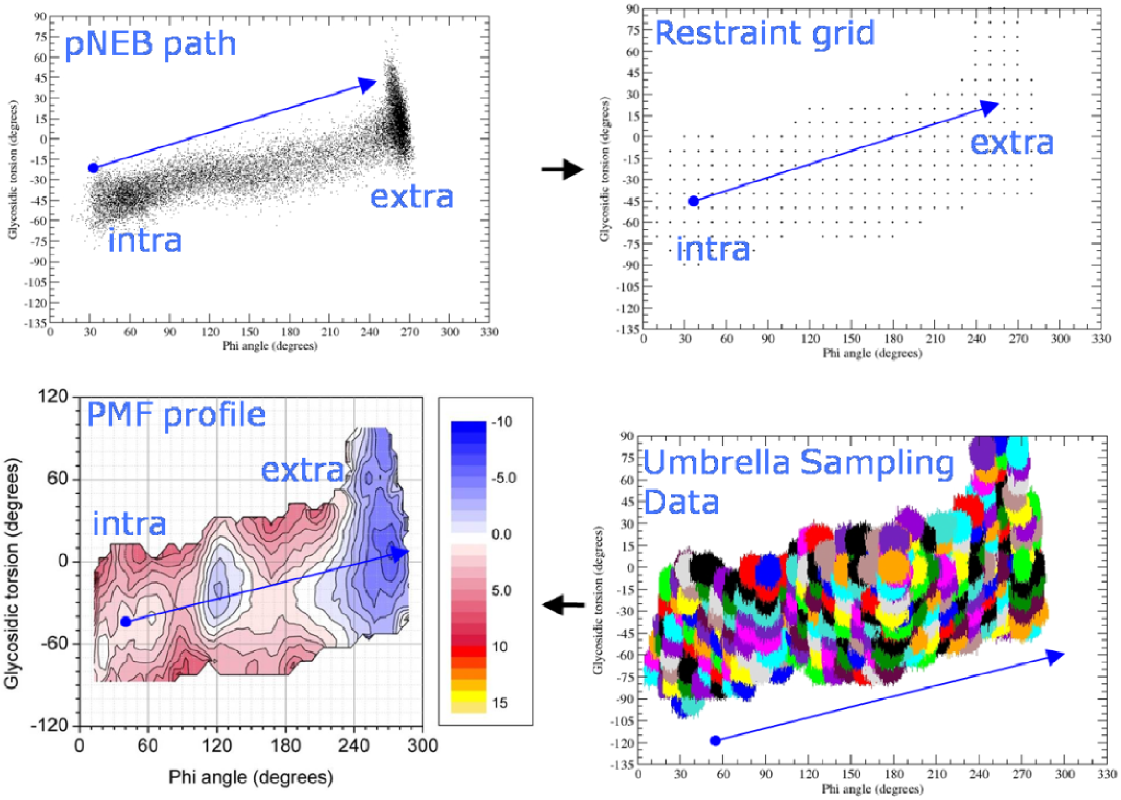


Figure 4-7. Workflow diagram for 8-oxo-G.wt through the major groove. All four figures are based on the two reaction coordinates; eversion and glycosidic torsion angle (x-axis and y-axis respectively). Last equilibration step of PNEB data (top left). Structures were extracted (top right) from the final step of PNEB data (top left) based on predefined grid points (top right). Each grid point (top right) represents the starting structure for our Umbrella Sampling simulations (bottom right). From the umbrella sampling simulations we were able to extract the two dimensional PMF profiles using 2D-WHAM (bottom left).

```
Eco-Fpg      PELPEVETSRRGIEPHLVGATILHA-VVRNGRLRWPVSEEIY--RLSDQPVLSVQRRAKY  57
Bst-Fpg      PELPEVETIRRTLLPLIVGKTIEDVRIFWPNIIRHPRDSEAFAARMIGQTVRGLERRGKF  60

Eco-Fpg      LLLELPEGWIIIHLGMSGSLRILPEELPPEKHDHVDLVMSNGKVLRYTDPRRFG-AWLWT 116
Bst-Fpg      LKFLLDRDALISHLRMEGRYAVASALEPLEPHTHVVFCFTDGSELRYRDVRKFGTMHVYA 120

Eco-Fpg      KELEGHN-VLTHLGPEPLSDDFNGEYLHQKCAKKKTAIKPWLMDNKLVVGVGNIYASESL 175
Bst-Fpg      KEEADRRPPLAELGPEPLSPAFSPAVLAERAVKTKRSVKALLLDQTVVAGFGNIYVDESL 180

Eco-Fpg      FAAGIHPDRLASSLSLAECELLARVIKAVLLRSIEQGGTTLKDFLQSDGKPGYFAQELQV 235
Bst-Fpg      FRAGILPGRPAASLSSKEIERLHEEMVATIGEAVMKGGSTVRTYVNTQGEAGTFQHHLYV 240

Eco-Fpg      YGRKGEPCRVCGTPIVATKHAQRATFYCRQCQK                            268
Bst-Fpg      YGRQGNPCKRCGTPIEKTVVAGRGTHYCPRCQR                            273
```

Figure 4-8. Alignment of *Eco*-Fpg and *Bst*-Fpg. Met73/76 and F110/113 are marked in bold and underlined. The alignment was performed using ClustalW2.

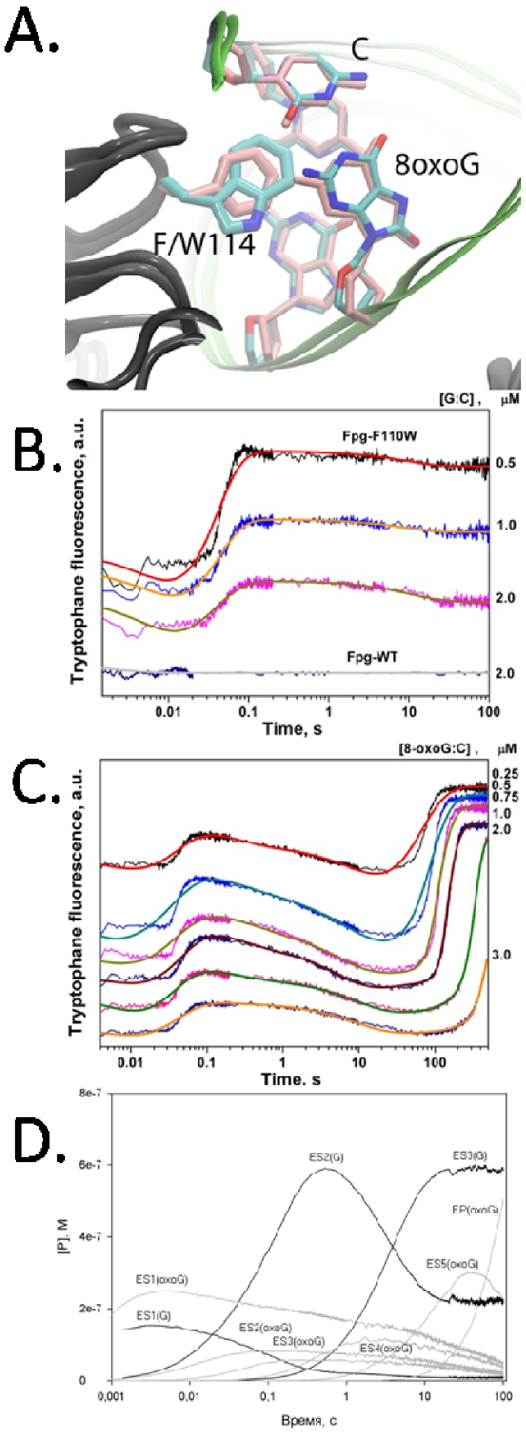
5′-CTCTCXCCTTCC-3′
3′-GAGAGCGGAAGG-5′

Figure 4-9. Oligonucleotide duplexes used in this work. X = 8-oxo-G or natural AP site.

4.3 Results and discussion

4.3.1 Fluorescence tracking of wedge insertion

To address whether wedge insertion indeed occurs with undamaged DNA and the sequence of events preceding and following the insertion, we have suggested that an F113W mutant of *Eco*-Fpg, in which the fluorescent Trp residue substituted for the wedge non-fluorescent Phe residue, may provide a suitable fluorescent reporter at the Fpg–DNA interface. To analyze possible effects of F113W mutation on the structure of the Fpg–DNA recognition complex, we first built a model for intrahelical 8-oxo-G using the structure of Fpg bound to intrahelical undamaged DNA (see methods section). We then introduced an *in silico* F113W wedge mutation in the corresponding position of the *Bst*-Fpg–DNA complex and performed 20 ns MD simulation for these complexes with native Phe and non-native Trp wedge. As shown in Figure 4-10, the replacement of the wild-type Phe phenyl wedge with the Trp indole has little effect on the DNA structure at the interrogation site, with the Trp wedge being inserted slightly deeper than that of Phe. Structural analyses were conducted on the simulated trajectories of the two systems (Figure 4-1). Relative to the crystal structure of Fpg cross-linked to undamaged DNA (PDB ID 2F5O), the RMSD of the DNA fluctuated near ~2 Å while that of the protein backbone remained near 1.2 Å. The similarities in the data between Fpg with either Phe or Trp wedge suggests that the mutation has little impact on the overall structure of the complex.


A.
C
8oxoG
F/W114
B.
[G:C] , μM
Fpg-F110W
0.5
1.0
2.0
Fpg-WT
2.0
Tryptophane fluorescence, a.u.
0.01
0.1
1
10
100
Time, s
C.
[8-oxoG:C] , μM
0.25
0.5
0.75
1.0
2.0
3.0
Tryptophane fluorescence, a.u.
0.01
0.1
1
10
100
Time. s
D.
8e-7
6e-7
4e-7
2e-7
0
[P], M
ES2(G)
ES3(G)
FP(oxoG)
ES5(oxoG)
ES1(oxoG)
ES1(G)
ES2(oxoG)
ES3(oxoG)
ES4(oxoG)
0,001
0,01
0,1
1
10
100
Время, с

E.)

Scheme I: THF-ligand

$$\mathrm{E + L} \underset{k_{-1}^{\mathrm{THF,Trp}}}{\overset{k_1^{\mathrm{THF,Trp}}}{\leftrightarrows}} \mathrm{EL1}^{\mathrm{THF,Trp}} \underset{k_{-2}^{\mathrm{THF,Trp}}}{\overset{k_2^{\mathrm{THF,Trp}}}{\leftrightarrows}} \mathrm{EL2}^{\mathrm{THF,Trp}} \underset{k_{-3}^{\mathrm{THF,Trp}}}{\overset{k_3^{\mathrm{THF,Trp}}}{\leftrightarrows}} \mathrm{EL3}^{\mathrm{THF,Trp}} \underset{k_{-4}^{\mathrm{THF,Trp}}}{\overset{k_4^{\mathrm{THF,Trp}}}{\leftrightarrows}} \mathrm{EL4}^{\mathrm{THF,Trp}}$$

Scheme II: AP-substrate

$$\mathrm{E + S} \underset{k_{-1}^{\mathrm{AP,Trp}}}{\overset{k_1^{\mathrm{AP,Trp}}}{\leftrightarrows}} \mathrm{ES1}^{\mathrm{AP,Trp}} \underset{k_{-2}^{\mathrm{AP,Trp}}}{\overset{k_2^{\mathrm{AP,Trp}}}{\leftrightarrows}} \mathrm{ES2}^{\mathrm{AP,Trp}} \underset{k_{-3}^{\mathrm{AP,Trp}}}{\overset{k_3^{\mathrm{AP,Trp}}}{\leftrightarrows}} \mathrm{ES3}^{\mathrm{AP,Trp}} \underset{k_{-4}^{\mathrm{AP,Trp}}}{\overset{k_4^{\mathrm{AP,Trp}}}{\leftrightarrows}} \mathrm{ES4}^{\mathrm{AP,Trp}} \xrightarrow{k_5^{\mathrm{AP,Trp}}} (\mathrm{EP_{gap}} \overset{K_{\mathrm{P}}^{\mathrm{Trp}}}{\leftrightarrows} \mathrm{E + P_{mix}})$$

Scheme III: 8-oxo-G-substrate

$$\mathrm{E + S} \underset{k_{-1}^{\mathrm{OG,Trp}}}{\overset{k_1^{\mathrm{OG,Trp}}}{\leftrightarrows}} \mathrm{ES1}^{\mathrm{OG,Trp}} \underset{k_{-2}^{\mathrm{OG,Trp}}}{\overset{k_2^{\mathrm{OG,Trp}}}{\leftrightarrows}} \mathrm{ES2}^{\mathrm{OG,Trp}} \underset{k_{-3}^{\mathrm{OG,Trp}}}{\overset{k_3^{\mathrm{OG,Trp}}}{\leftrightarrows}} \mathrm{ES3}^{\mathrm{OG,Trp}} \underset{k_{-4}^{\mathrm{OG,Trp}}}{\overset{k_4^{\mathrm{OG,Trp}}}{\leftrightarrows}} \mathrm{ES4}^{\mathrm{OG,Trp}} \underset{k_{-5}^{\mathrm{OG,Trp}}}{\overset{k_5^{\mathrm{OG,Trp}}}{\leftrightarrows}} \mathrm{ES5}^{\mathrm{OG,Trp}} \xrightarrow{k_6^{\mathrm{OG,Trp}}} (\mathrm{EP_{gap}} \overset{K_{\mathrm{P}}^{\mathrm{Trp}}}{\leftrightarrows} \mathrm{E + P_{mix}})$$

Scheme IV: G-ligand (Fpg-F110W only)

$$\mathrm{E + L} \underset{k_{-1}^{\mathrm{G,Trp}}}{\overset{k_1^{\mathrm{G,Trp}}}{\leftrightarrows}} \mathrm{EL1}^{\mathrm{G,Trp}} \underset{k_{-2}^{\mathrm{G,Trp}}}{\overset{k_2^{\mathrm{G,Trp}}}{\leftrightarrows}} \mathrm{EL2}^{\mathrm{G,Trp}} \underset{k_{-3}^{\mathrm{G,Trp}}}{\overset{k_3^{\mathrm{G,Trp}}}{\leftrightarrows}} \mathrm{EL3}^{\mathrm{G,Trp}}$$

Figure 4-10. **A**.) Overlap of structures from simulation of *Bst*-Fpg–DNA complex with a wild-type Phe113 wedge (pink) and a mutant Trp113 wedge (colored by atom). Structure is shown only for the wedge and base pairs at the insertion site. **B**.) Experimental (jagged traces) and fit (smooth curves) time courses of Trp fluorescence changes during cleavage of the 8-oxo-G:C substrate by *Eco*-Fpg-F113W. **C**.) Experimental (jagged traces) and fit (smooth curves) time courses of Trp fluorescence changes during binding of *Eco*-Fpg-F110W and wild-type *Eco*-Fpg to undamaged DNA. A.u., arbitrary units. **D**.) Kinetic simulation of accumulation and disappearance of various enzyme–substrate complexes during the interaction of *Eco*-Fpg-F113W with undamaged (G-ligand) and damaged (8-oxo-G substrate) DNA. **E**.) Kinetic schemes of *Eco*-Fpg and *Eco*-Fpg-F113W interactions with the G-ligand, THF-ligand, AP-substrate and 8-oxo-G substrate as detected by Trp fluorescence. E, enzyme; L, uncleavable ligand; S, substrate; EL$i^{\mathrm{X,Trp}}$ or ES$i^{\mathrm{X,Trp}}$, various fluorescently discernible complexes of X (substrate or ligand) with Fpg; $\mathrm{P_{gap}}$, gapped double-stranded oligonucleotide reaction product; $\mathrm{P_{mix}}$, the $\mathrm{P_{gap}}$ product dissociated into three separate strands.

Having determined that placing Trp instead of Phe into the Fpg–DNA complex should not have an adverse effect on its structure, we have constructed and purified the respective *Eco*-Fpg mutant and investigated the steady-state and stopped-flow kinetics of its reaction with undamaged DNA, as well as with DNA containing an uncleavable abasic tetrahydrofuran moiety (THF), natural abasic site (AP) or 8-oxo-G opposite to C. Steady-state K_M and k_{cat} of *Eco*-Fpg-F113W cleaving the 8-oxo-G-substrate were similar to those of *Eco*-Fpg (Table 4-5), supporting the results of the modeling. When rate constants associated with pre-catalytic conformational adjustment and catalysis were measured by stopped-flow kinetics using the THF-ligand, AP-, or 8-oxo-G-substrate (Figure 4-10B), the absolute values for different steps varied (Table 4-4) but the minimal reaction schemes remained the same as for wild-type *Eco*-Fpg (Schemes I–III, Figure 4-10E). Notably, the most pronounced changes in the rate constants occurred at the stages presumed to involve substrate destabilization, wedge insertion and base eversion, the processes which are likely to be kinetically affected by the F113W mutation; the steps of primary binding (rate constants k_1 and k_{-1}), active site isomerization (k_5, k_{-5}) and catalysis (k_6) were influenced to a much less extent, as expected for stages presumably not involving wedge movement. The most drastic change was observed with the G-ligand, binding of which to wild-type *Eco*-Fpg could be described by a fast single-step equilibrium process (characteristic time $\square_1 \sim 5$ ms), but the association of *Eco*-Fpg-F113W with this ligand could be separated into three fluorescently discernible stages, two of which clearly followed the fast binding ($\square_2 \sim 50$ ms, $\square_3 \sim 10$ s) (Scheme IV, Fig. Figure 4-10E, Table 4-4). These changes in fluorescence most likely reflect conformational changes not observed in wild-type *Eco*-Fpg due to lack of an appropriate fluorescent reporter at the DNA-binding interface.

Table 4-4. Pre-steady-state parameters of *Eco*-Fpg WT and mutants; Trp fluorescence

	Eco-Fpg-WT[a]				*Eco*-Fpg-F110W[b]				g-F113A[b]
	G	THF[c]	AP[d]	oxoG[e]	G[f]	THF[c]	AP[d]	oxoG[e]	oxoG[g]
k_1, M^{-1}s^{-1}	(230±40) ×10^6	50±20)×10^6	(800±30) ×10^6	(320±15) ×10^6	(260±10) ×10^6	(140±6) ×10^6	(230±8) ×10^6	(340±10) ×10^6	(100±14) ×10^6
k_{-1}, s^{-1}	2700 ±200	270 ±15	250 ±20	890 ±25	1100 ±90	920 ±30	300 ±40	700 ±25	990 ±250
k_2, s^{-1}		6.0 ±0.2	36 ±3	250 ±18	50.6 ±4.8	11.9 ±1.3	13.1 ±1.1	27.7 ±1.4	54 ±13
k_{-2}, s^{-1}		0.02 ±0.005	65 ±8	2.4 ±0.4	1.8 ±0.4	31.6 ±1.8	39.4 ±5.2	63 ±2	40 ±3
k_3, s^{-1}		10.0 ±0.6	10.0 ±0.5	6.7 ±0.8	0.24 ±0.05	7.1 ±0.5	3.4 ±0.1	22 ±1	
k_{-3}, s^{-1}		0.6 ±0.1	40 ±4	46 ±0.6	0.09 ±0.01	1.71 ±0.16	1.74 ±0.08	28 ±1	
k_4, s^{-1}		0.04 ±0.007	11.0 ±0.8	9.1 ±0.6		0.26 ±0.08	0.52 ±0.01	4.5 ±0.1	
k_{-4}, s^{-1}		0.01 ±0.001	1.0 ±0.1	2.4 ±0.2		0.22 ±0.04	0.03 ±0.01	1.9 ±0.1	
k_5, s^{-1}				0.20 ±0.01				0.20 ±0.01	
k_{-5}, s^{-1}				0.03 ±0.007				0.02 ±0.002	
k_6, s^{-1}			0.20±0.01	0.04 ±0.006			0.30±0.04	0.021 ±0.004	
K_P, M			8±0.3)×10^{-6}	0±0.3)×10^{-6}			±0.1)×10^{-6}	6±0.6)×10^{-6}	

Estimate ± standard error of the estimate.

[a] From $REF.

[b] This work.

[c] Scheme I:

$$\mathrm{E + L} \underset{k_{-1}^{\mathrm{THF,Trp}}}{\overset{k_1^{\mathrm{THF,Trp}}}{\rightleftarrows}} \mathrm{EL1^{THF,Trp}} \underset{k_{-2}^{\mathrm{THF,Trp}}}{\overset{k_2^{\mathrm{THF,Trp}}}{\rightleftarrows}} \mathrm{EL2^{THF,Trp}} \underset{k_{-3}^{\mathrm{THF,Trp}}}{\overset{k_3^{\mathrm{THF,Trp}}}{\rightleftarrows}} \mathrm{EL3^{THF,Trp}} \underset{k_{-4}^{\mathrm{THF,Trp}}}{\overset{k_4^{\mathrm{THF,Trp}}}{\rightleftarrows}} \mathrm{EL4^{THF,Trp}}$$

[d] Scheme II:

$$\mathrm{E + S} \underset{k_{-1}^{\mathrm{AP,Trp}}}{\overset{k_1^{\mathrm{AP,Trp}}}{\rightleftarrows}} \mathrm{ES1^{AP,Trp}} \underset{k_{-2}^{\mathrm{AP,Trp}}}{\overset{k_2^{\mathrm{AP,Trp}}}{\rightleftarrows}} \mathrm{ES2^{AP,Trp}} \underset{k_{-3}^{\mathrm{AP,Trp}}}{\overset{k_3^{\mathrm{AP,Trp}}}{\rightleftarrows}} \mathrm{ES3^{AP,Trp}} \underset{k_{-4}^{\mathrm{AP,Trp}}}{\overset{k_4^{\mathrm{AP,Trp}}}{\rightleftarrows}} \mathrm{ES4^{AP,Trp}} \xrightarrow{k_5^{\mathrm{AP,Trp}}} (\mathrm{EP_{gap}} \overset{K_{\mathrm{P}}^{\mathrm{Trp}}}{\rightleftarrows} \mathrm{EP_{mix}})$$

[e] Scheme III:

$$\mathrm{E + S} \underset{k_{-1}^{\mathrm{oxoG,Trp}}}{\overset{k_1^{\mathrm{oxoG,Trp}}}{\rightleftarrows}} \mathrm{ES1^{oxoG,Trp}} \underset{k_{-2}^{\mathrm{oxoG,Trp}}}{\overset{k_2^{\mathrm{oxoG,Trp}}}{\rightleftarrows}} \mathrm{ES2^{oxoG,Trp}} \underset{k_{-3}^{\mathrm{oxoG,Trp}}}{\overset{k_3^{\mathrm{oxoG,Trp}}}{\rightleftarrows}} \mathrm{ES3^{oxoG,Trp}} \underset{k_{-4}^{\mathrm{oxoG,Trp}}}{\overset{k_4^{\mathrm{oxoG,Trp}}}{\rightleftarrows}} \mathrm{ES4^{oxoG,Trp}} \underset{k_{-5}^{\mathrm{oxoG,Trp}}}{\overset{k_5^{\mathrm{oxoG,Trp}}}{\rightleftarrows}} \mathrm{ES5^{oxoG,Trp}} \xrightarrow{k_6^{\mathrm{oxoG,Trp}}} (\mathrm{EP_{gap}} \overset{K_{\mathrm{P}}^{\mathrm{Trp}}}{\rightleftarrows} \mathrm{EP_{mix}})$$

[f] Scheme IV:

$$\mathrm{E + L} \underset{k_{-1}^{\mathrm{G,Trp}}}{\overset{k_1^{\mathrm{G,Trp}}}{\rightleftarrows}} \mathrm{EL1^{G,Trp}} \underset{k_{-2}^{\mathrm{G,Trp}}}{\overset{k_2^{\mathrm{G,Trp}}}{\rightleftarrows}} \mathrm{EL2^{G,Trp}} \underset{k_{-3}^{\mathrm{G,Trp}}}{\overset{k_3^{\mathrm{G,Trp}}}{\rightleftarrows}} \mathrm{EL3^{G,Trp}}$$

[g] Scheme V:

$$\mathrm{E' + S} \underset{k_{-1}^{\mathrm{oxoG,Trp}}}{\overset{k_1^{\mathrm{oxoG,Trp}}}{\rightleftarrows}} \mathrm{E'S1^{oxoG,Trp}} \underset{k_{-2}^{\mathrm{oxoG,Trp}}}{\overset{k_2^{\mathrm{oxoG,Trp}}}{\rightleftarrows}} \mathrm{E'S2^{oxoG,Trp}}$$

In all Schemes: E, enzyme; E’, inactive enzyme; L, uncleavable ligand; S, substrate; EL$i^{\mathrm{X,Trp}}$ or ES$i^{\mathrm{X,Trp}}$, various fluorescently discernible complexes of X (substrate or ligand) with Fpg; P$_{\mathrm{gap}}$, gapped double-stranded oligonucleotide reaction product; P_{mix} is P_{gap} dissociated into three separate strands.

Table 4-5. Pre-steady-state parameters of *Eco*-Fpg-WT and *Eco*-Fpg-F113A; aPu fluorescence

	Eco-Fpg-WT[a]	*Eco*-Fpg-F113A[a]
k_1, $M^{-1}\times s^{-1}$	$(9.7 \pm 0.4)\times 10^5$	$(3.6 \pm 0.5)\times 10^5$
k_{-1}, s^{-1}	1.7 ± 0.1	2.6 ± 0.6
k_2, s^{-1}	0.87 ± 0.07	0.50 ± 0.06
K_P, M	$(4.4 \pm 0.6)\times 10^{-6}$	$(3.3 \pm 0.6)\times 10^{-6}$

Estimate ± standard error of the estimate.

[a] Scheme VI:

$$\mathrm{E} + \mathrm{S} \underset{k_{-1}^{\mathrm{aPu}}}{\overset{k_1^{\mathrm{aPu}}}{\rightleftarrows}} \mathrm{ES}^{\mathrm{aPu}} \xrightarrow{k_2^{\mathrm{aPu}}} (\mathrm{EP}_{\mathrm{gap}} \overset{K_{\mathrm{P}}^{\mathrm{aPu}}}{\rightleftarrows} \mathrm{EP}_{\mathrm{mix}})$$

The species in the Scheme are named in the same way as in Table 4-4.

A comparison of the appearance and disappearance of different species during binding of the G-ligand and 8-oxo-G-substrate by Fpg-F113W (Figure 4-10D) suggests that the ES1 species in both Schemes III and IV are the same and characterize the primary enzyme–DNA binding. ES2 in Scheme IV apparently corresponds to ES2 or ES3 in Scheme III, the steps that reflect DNA destabilization, Phe intercalation, and possibly 8-oxo-G eversion. Finally, the ES3 species in Scheme IV may either correspond to ES4 or ES5 species in Scheme III, or reflect a conformational change step absent in the productive substrate encounter. In any case, sampling of the normal G base by *Eco*-Fpg-F113W seems to be complicated and involve several conformational changes, most likely not stopping after insertion of Phe113 inferred from the structure of *Bst*-Fpg complexed with undamaged DNA.[71] As the reaction schemes for other substrates and ligands processed by wild-type *Eco*-Fpg and *Eco*-Fpg-F113W are similar, one may expect that wild-type *Eco*-Fpg also deeply samples the G base. Structural studies with two other repair enzymes, human 8-oxo-Guanine-DNA glycosylase hOGG1[165,166] and uracil-DNA glycosylase[167] suggest that sampling of normal bases, going to a considerable depth but not all the way to the Michaelis complex, may be a common theme in discrimination of normal bases from damaged ones. Effects of wedge mutations on other glycosylases have been studied by steady-state kinetics[168-170]; however, only for uracil-DNA glycosylase it was shown that the mutation affects a defined conformational transition in a multistep reaction coordinate.[171]

4.3.2 Computational analysis of the role of the wedge in lesion processing

These observations suggests that the wedge insertion by Fpg could be a critical step in initiating the eversion of both undamaged and damaged deoxynucleoside (no pre-eversion structure is available for the latter), and that abolishment of this residue should make the eversion unfavorable. In the series of crystal structures of undamaged DNA bound to Fpg, the aromatic wedge causes significant buckling at the intercalation site, potentially probing the stability of the interrogated base pair.[71] In order to understand the effect of the wedge on lesion processing, we performed free energy calculations for 8OG eversion for WT Fpg as well as for the F113W mutant. We used nudged elastic band (NEB) calculations to map out the pathway of eversion for 8OG through the major groove and into the enzyme active site. The free energy profile along this pathway was calculated using umbrella sampling simulations on a grid surrounding the region sampled by the NEB pathway. Two progress variables were selected as coordinates for the free energy change: the eversion angle of the base and the rotation of the glycosidic angle of the everting base (which changes from *anti* to *syn* in the intrahelical and extrahelical crystallographic structures). The resulting free energy profiles are shown in Figure 4-11.

Eversion of 8-oxo-G in the WT enzyme is a predominantly downhill process in the simulations, encountering barriers of only a few kcal/mol. In contrast, the barrier to breaking the 8-oxo-G:C base pair during initial eversion (approaching 60°) is nearly 15 kcal/mol, much higher than for the WT Fpg. Since single molecule fluorescence experiments estimate the energetic barrier to 1D diffusion of Fpg along DNA at only ~2 kcal/mol, it seems unlikely that the residence time of the enzyme at a particular interrogation site would be sufficient to overcome an eversion barrier that is more than 10 kcal/mol higher than that of translocation.

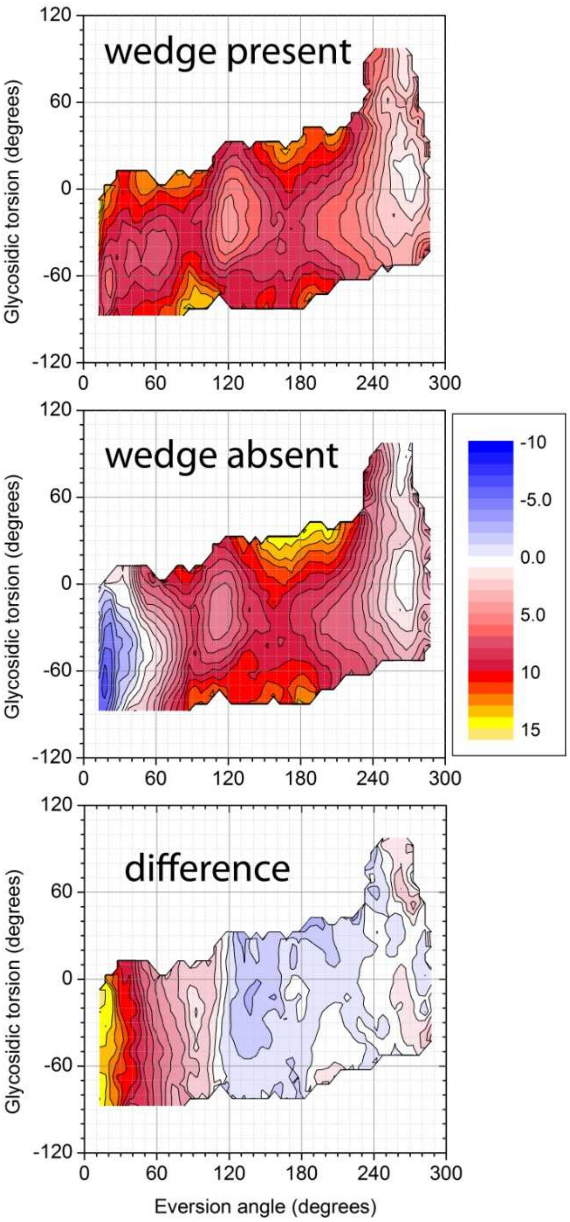
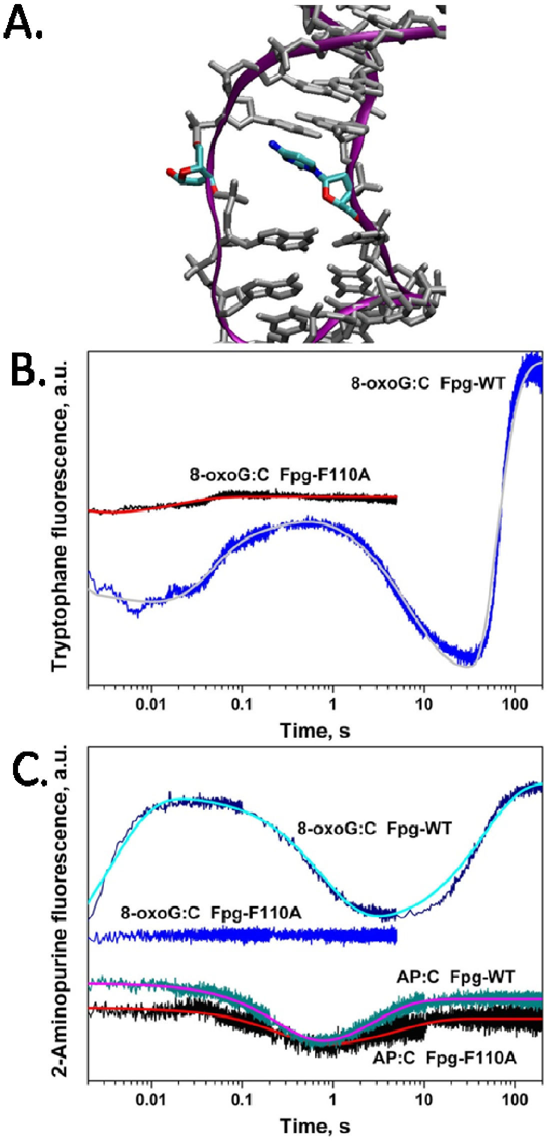


Figure 4-11. Comparison of free energy profiles for lesion eversion in the Fpg–DNA complex with Phe (**wedge present**) and Ala (**wedge absent**) in the wedge position. The intrahelical position is on the left, and extrahelical (active site) is on the right. The major difference between the profiles is that the intrahelical state is significantly stabilized in the F113A mutant, as demonstrated in (**difference**), which shows the difference (WT-mutant) in the free energy surfaces.

4.3.3 Wedge deletion mutant lacks glycosylase function but retains AP-lyase activity

Therefore, based on the results of our computational modeling, an Fpg wedge mutant should be less competent at the stage of lesion eversion. We constructed the *Eco*-Fpg-F113A mutant and analyzed its activity on 8-oxo-G- and AP-substrates using steady-state and stopped-flow kinetics. Under steady-state conditions, the mutant enzyme was unable to process 8-oxo-G-substrates but maintained some residual activity on AP-substrates (Table 4-5), in line with observations that many mutations of critical residues in Fpg/Nei enzymes fully abolish their DNA glycosylase but not AP lyase activity.[170,172-174] In order to understand the residual activity for the AP-substrate, we performed a simulation in solution for duplex DNA of the same sequence as used for the complex, but with an AP site replacing the 8-oxo-G. We observed that the AP substrate preferred the extrahelical position even in the absence of Fpg, due to rearrangement of the backbone to minimize the gap resulting from the absence of a base (Figure 4-12A). The eversion angle for the AP-substrate was stable near 200° (Figure 4-12); thus the model indicates that no wedge would be required to facilitate eversion of the AP-substrate.



D.

Scheme V:

$$E + S \underset{k_{-1}^{2\text{-aPu}}}{\overset{k_1^{2\text{-aPu}}}{\leftrightarrows}} ES^{2\text{-aPu}} \xrightarrow{k_2^{2\text{-aPu}}} (EP_{gap} \overset{K_P^{2\text{-aPu}}}{\leftrightarrows} E + P_{mix})$$

Figure 4-12. **A**, Structure of duplex DNA in solution containing AP-substrate that has spontaneously everted. Only heavy atoms of the DNA are shown. **B, C**, Experimental (jagged traces) and fit (smooth curves) time courses of Trp (**B**) or 2-aPu (**C**) fluorescence changes during cleavage of the 8-oxo-G:C (**B, C**) or AP substrates (**C**) by wild-type *Eco*-Fpg and *Eco*-Fpg-F113A. A.u., arbitrary units. **D**, kinetic schemes of *Eco*-Fpg and *Eco*-Fpg-F113A interactions with the AP- and 8-oxo-G-substrates as detected by 2-aPu fluorescence. $ES^{2\text{-aPu}}$, Fpg–DNA complex; other designations are the same as in Figure 4-10E.

When analyzed by the stopped-flow technique, the interaction of *Eco*-Fpg-F113A with the 8-oxo-G-substrate could be described by two stages with kinetic parameters similar to the two initial steps observed for wild-type *Eco*-Fpg (Table 4-5), and the reaction did not proceed any further, as could be followed until the Trp bleaching made the detection impossible at >500 s (Figure 4-12B and data not shown). Interestingly, when a fluorescent 2-aminopurine (2-aPu) reporter was incorporated 5' or 3' to the 8-oxo-G residue, wild-type *Eco*-Fpg showed at least two conformational transitions preceding the catalytic step, while *Eco*-Fpg-F113A did not produce any change (Figure 4-12C). As changes in 2-aPu fluorescence reflect changes in the hydrophobicity of its environment [175], it can be inferred that no appreciable conformational transitions occur in DNA in the vicinity of the damaged base when *Eco*-Fpg-F113A binds its substrate. Consistent with the computational modeling, we conclude that *Eco*-Fpg-F113A forms a normal primary encounter complex with the 8-oxo-G-substrate but failure to insert a bulky wedge next to the damaged base precludes further reaction steps. Reaction of *Eco*-Fpg-F113A with the AP-substrate could also be followed by Trp fluorescence but a low amplitude of the absolute changes in fluorescence intensity made the signal-to-noise ratio low and quantification unreliable (not shown); hence, we have analyzed the reaction coordinate by 2-aPu fluorescence and obtained the values for rate constants of individual conformational transitions that were similar for wild-type *Eco*-Fpg and *Eco*-Fpg-F113A (Figure 4-12C, Table 4-5). In this case, the changes in 2-aPu fluorescence are likely due to intercalation of other plugging residues (Met76 and Arg111). The stopped-flow results agree with the steady-state activity of Eco-Fpg-F110A and with the modeling that indicated relaxed requirements for wedge insertion for AP site cleavage.

To analyze the effect of the wedge mutation on the ability of Fpg to bend substrate DNA, we have followed the efficiency of Förster resonance energy transfer (FRET) in the stopped-flow mode. The 8-oxo-G-substrate was modified at the opposite 5'-termini with either a pair of fluorescence dyes (Cy3 and Cy5), in which case the emission from Cy5 after photon transfer from Cy3 was monitored, or with a dye–quencher pair (fluorescein and Dabcyl, respectively) to observe quenching of fluorescein emission. When the Cy3/Cy5 8-oxo-G-substrate was cleaved by wild-type *Eco*-Fpg, the Cy5 fluorescence initially slightly increased and then started to decay in several discernible steps (

Figure 4-13A,B); the fluorescein emission from the fluorescein/Dabcyl 8-oxo-G-substrate almost exactly inversely mirrored the Cy5 emission (

Figure 4-13A). The data were found to be optimally fit by SchemeVI, and the obtained kinetic constants are presented in Table 4-6. Importantly, the reaction scheme shows that the bending angle of substrate DNA changes both during the pre-catalytic adjustment of the enzyme–substrate conformation and during the catalytic step. The enzyme essentially over-bends DNA during primary binding and then relaxes it in two steps. However, with the *Eco*-Fpg-F113A mutant, the first phase of the Cy5 fluorescence decay was significantly prolonged (

Figure 4-13B), suggesting that both forms of the enzyme quickly bend DNA but *Eco*-Fpg-F113A is much slower to proceed further. It can be seen that *Eco*-Fpg-F113A ultimately cleaves the substrate at >1,000 s but, after the bending step, the rate constants for this enzyme form are ~1–2 orders of magnitude lower than for the wild-type enzyme.

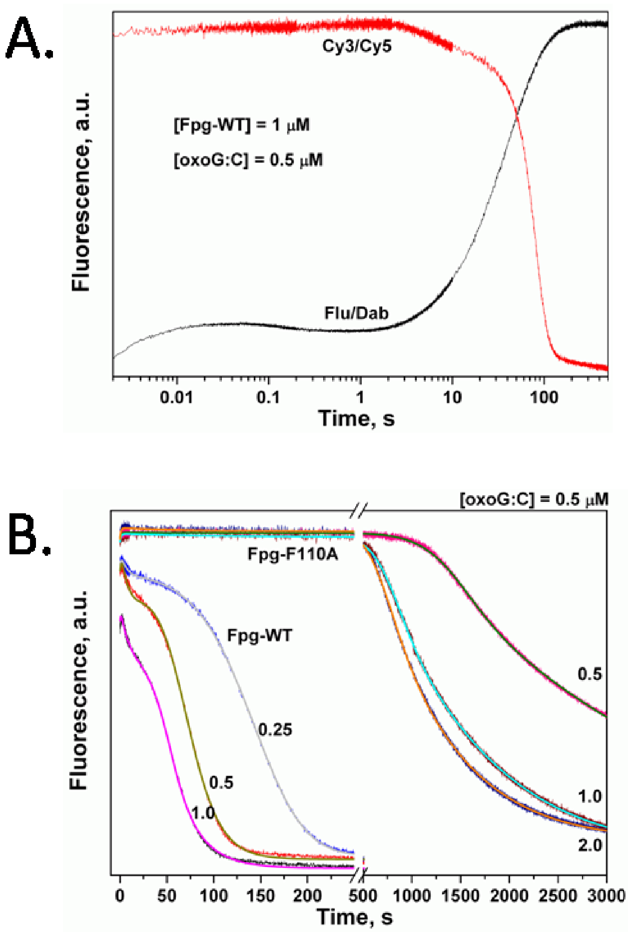


Figure 4-13. Time courses of FRET efficiency during cleavage of 8-oxo-G-substrate by Fpg. **A**.) a comparison of Cy3/Cy5 and fluorescein/Dabcyl (Flu/Dab) FRET time courses for wild-type *Eco*-Fpg. **B**.) FRET time courses for wild-type *Eco*-Fpg and *Eco*-Fpg-F113A. Jagged traces represent experimental data, smooth curves show the fit to Scheme VI. A.u., arbitrary units.

Table 4-6. Pre-steady-state parameters of *Eco*-Fpg-WT and *Eco*-Fpg-F110A; Cy3/Cy5 FRET

	Eco-Fpg-WT	*Eco*-Fpg-F110A
k_1, $\mathrm{M^{-1}\times s^{-1}}$	$(0.59 \pm 0.14)\times 10^6$	$(0.3 \pm 0.05)\times 10^6$
k_{-1}, $\mathrm{s^{-1}}$	0.83 ± 0.08	0.4 ± 0.06
k_2, $\mathrm{s^{-1}}$	0.15 ± 0.04	0.0018 ± 0.0002
k_{-2}, $\mathrm{s^{-1}}$	0.0020 ± 0.0004	3.5E-4 ± 2.7E-4
k_3, $\mathrm{s^{-1}}$	0.046 ± 0.008	0.0075 ± 0.0034
K_P, M	$(0.23 \pm 0.11)\times 10^6$	$(0.029 \pm 0.006)\times 10^6$

Estimate ± standard error of the estimate.

[a]Scheme VII:

$$\mathrm{E} + \mathrm{S} \underset{k_{-1}^{\mathrm{Cy}}}{\overset{k_1^{\mathrm{Cy}}}{\rightleftarrows}} \mathrm{ES1}^{\mathrm{Cy}} \underset{k_{-2}^{\mathrm{Cy}}}{\overset{k_2^{\mathrm{Cy}}}{\rightleftarrows}} \mathrm{ES2}^{\mathrm{Cy}} \xrightarrow{k_3^{\mathrm{Cy}}} (\mathrm{EP}_{\mathrm{gap}} \overset{K_\mathrm{P}^{\mathrm{Cy}}}{\rightleftarrows} \mathrm{EP}_{\mathrm{mix}})$$

The species in the Scheme are named in the same way as in Table 1.

4.3.4 How does the wedge destabilize the interrogated base pair?

Since the experimental data confirmed the role of the wedge as suggested by the calculated free energy profiles for lesion eversion, we performed further analyses on the computational data in order to gain insight into the mechanism of lesion destabilization by the wedge. As discussed above, the interrogated base pair shows significant buckling in the crystal structure of Fpg interrogating an intrahelical G. Our simulations reproduce this buckle for WT Fpg interrogating 8-oxo-G, but the buckle is absent with the F113A Fpg (Figure 4-14). In order to understand the broader effect of buckle during lesion eversion, we measured the average buckle angle for structures in each bin on the 2D grid corresponding to the eversion free energy profile (Figure 4-3). The data confirm that the wedge presence significantly increases 8-oxo-G:C buckle across the entire range of structures prior to breaking of the base pair.

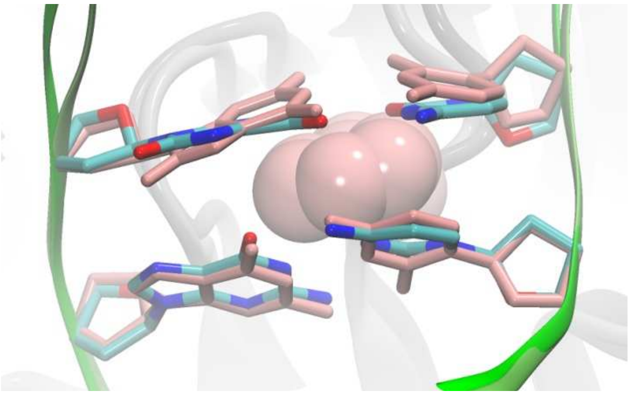


Figure 4-14. Structure overlap for the interrogation site with F113A mutant (colored by atom) and with WT Phe wedge (pink).

What is the energetic consequence of increased buckle? We measured the length of each of the three 8-oxo-G:C Watson-Crick hydrogen bonds over the entire eversion profile (Figure 4-4) and observed that all 3 hydrogen bonds break at an earlier extent of eversion when the wedge is present. Analogously, at a given extent of eversion, the hydrogen bonds are longer with the wedge present, and thus presumably weaker. This wedge-induced weakening of hydrogen bonds may explain the increased difficulty of F113A in processing the 8-oxo-G lesion.

In order to confirm this impact on base pair strength, and investigate other possible sources of the differences in eversion free energy profiles seen in Figure 2C, we performed a pair-wise decomposition of the total potential energy for structures sampled on the energy landscape. For each bin on the 2D grid, we calculated the average interaction energy between each pair of residues as well as the internal energy for each residue. Examination of the resulting data revealed that 4 key interactions significantly differ with the wedge present or absent (Figure 4-15). These 4 main effects of the wedge include weakening the 8-oxo-G:C base pair (as expected from the hydrogen bond analysis), reducing the stacking of the buckled 8-oxo-G against the 3’ G (Figure 4-15), reduction of the internal strain in the 8-oxo-G base, and an additional interaction with Met76 discussed below. Plots for the 4 individual interactions are shown in Figure 4-16.

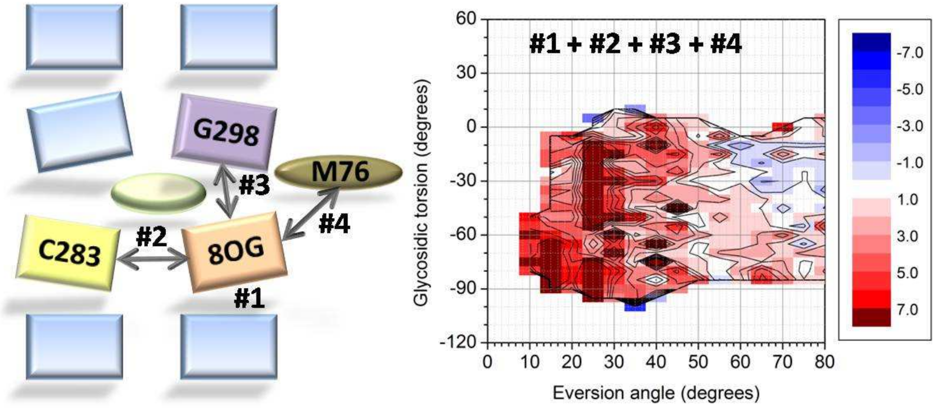


Figure 4-15. Color-coded difference in interaction energies due to the wedge, mapped onto the initial stages of the 2D eversion profile in Figure 4-11C. The figure shows the sum of four important interactions described in the text. Positive values reflect more favorable interactions in the mutant that increase the energy required for lesion eversion.

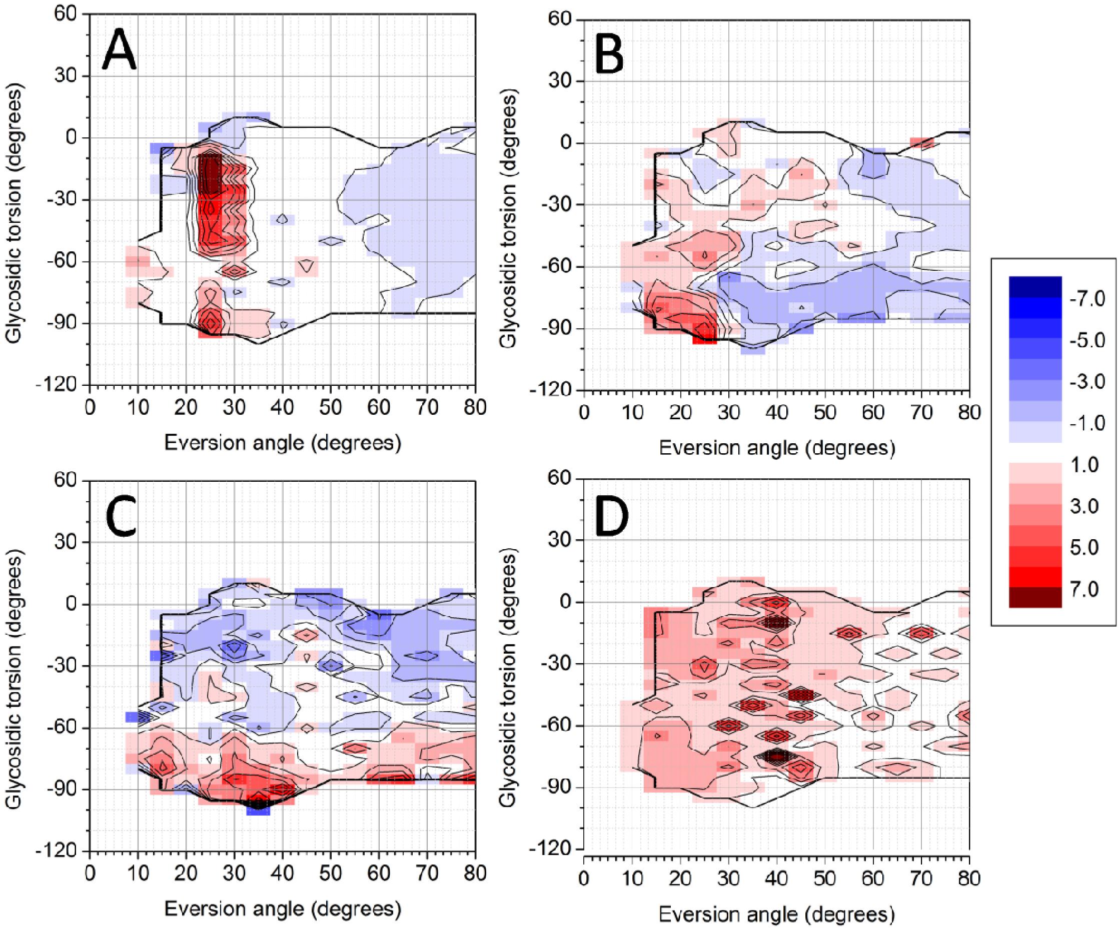


Figure 4-16. **A**.) Energy decomposition differences for 8-oxo-G:C base pair **B**.) 8-oxo-G stacking against the 3' G **C**.) internal strain in the 8-oxo-G **D**.) and the interaction between 8-oxo-G and Met76. Values are shown as *Bst*-Fpg-WT minus *Bst*-Fpg-F113A, thus positive numbers reflect increased stability in the mutant complex.

While 3 of the 4 key effects of the wedge arise directly from the 8-oxo-G buckling, the last involves Met76, a completely conserved residue in Fpg. In the intrahelical state, Met76 is at the end of a strained loop region that undergoes a conformational change upon eversion, resulting in motion of Met76 into the minor groove and filling of the gap left by the everted lesion. In the absence of the bulky wedge in F113A Fpg, the intrahelical state undergoes steric repacking that allows the loop to relax prior to eversion, with the Met76 side chain filling the space left by the missing wedgeand forming a favorable interaction with 8-oxo-G (Figure 4-17). The additional space also results in less effective gap filling by Met76 in the F113A extrahelical state. The overall effect is a significant stabilization of the intrahelical 8-oxo-G (Figure 4-16D), contributing to the inability of F113A Fpg to evert and process the lesion.

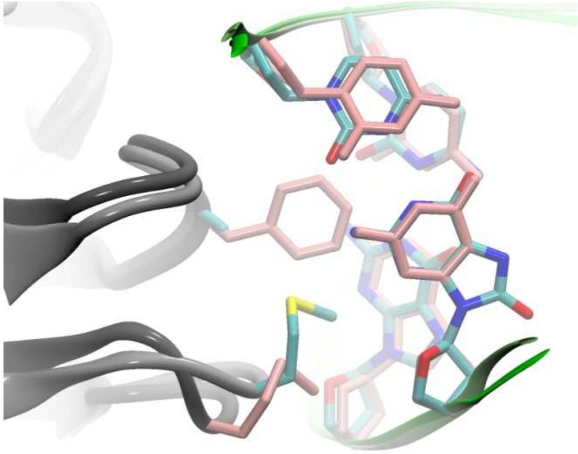


Figure 4-17. Comparison of structures at the intrahelical interrogation site for F113A (colored by atom) and WT (pink). Loss of the phenyl ring wedge allows Met76 to relax prior to eversion, in contrast to WT Fpg where relaxation results in filling the gap left by the everted lesion.

4.4 Conclusions

Therefore, we have combined high-quality molecular dynamics and stopped-flow kinetics with fluorescence detection to analyze a critical conformational change in the reaction catalyzed by formamidopyrimidine-DNA glycosylase. Our results suggest that, during lesion search, Fpg likely samples undamaged bases by wedge insertion and several conformational changes in the enzyme–DNA complex. A similar scheme was recently proposed for uracil-DNA glycosylase based on structural data.[167] The strained conformation of kinked DNA after Fpg binding may provide a source of energy for catalysis; however, the productive reaction must be triggered by wedge insertion, which therefore represents a critical step in the lesion recognition process. The wedge actively destabilizes the lesion through a buckling process that weakens the Watson-Crick hydrogen bonds, reduces 3’ stacking, and increases internal strain in the lesion. This mechanism could be common for many DNA glycosylases that possess an aromatic intercalating wedge.[173,176-178]

Chapter 5 Exo-site lesion discrimination in the Fpg glycosylase

Abstract

Fpg/MutM is a bacterial DNA glycosylase known to excise the oxidative product 8-oxo-guanine (oxoG) when it is extra-helical in the active site suggesting that the nucleobase must first evert out from the intra-helical state.[70,71] It is currently unclear whether base eversion occurs for guanine (G) and if it does to what extent. Little is known about the mechanism by which Fpg/MutM discriminates between G and oxoG and whether this function is conserved in other glycosylases. It is also unclear during base eversion whether the everting nucleobase leaves the intra-helical state through the minor or major groove. We hypothesize that Fpg/MutM facilitates Watson-Crick base pair opening of G:C and oxoG:C after which the enzyme can distinguish between G and oxoG and promote complete eversion of oxoG into the active site for excision while inhibiting further eversion of G. To measure whether eversion occurs for both nucleobases when bound to Fpg/MutM, a series of time course fluorescence experiments were used and show that both of these nucleobases adopt several states outside the intra-helical when complexed to Fpg/MutM suggesting that base eversion occurs for both. To understand whether base eversion occurs through either the minor or major groove, several computational techniques were used to obtain energetic profiles to eversion of G and oxoG through both grooves.[91,94-96] From these energetic profiles, the major groove direction was found to be the more energetically favorable pathway for both G and oxoG. Upon initial base pair opening, it was found that Arg 263 and Lys

257 aid in discriminating the two nucleobases by a repositioning of the Zinc finger. The resulting major groove energy profiles along with the fluorescence data suggest a transiently stable exo-site for both systems. We hypothesize that specific residue to residue interactions were occurring which contributed to the initial breaking of the Watson-Crick hydrogen bonds and the stabilization of each nucleobase in the exo-site. To understand these specific interactions, structural analysis and pair-wise energy decomposition were employed. Specific interactions that discriminate between the two nucleobases at the exo-site were discovered. Two specific interactions stabilize oxoG at the exo-site: the 5' phosphate group two nucleobases adjacent to oxoG, and the conserved Asn173. The interactions that occur with oxoG result in a lowering of the energetic pathway to eversion. In this work we investigate how the events in initial base pair opening coupled with the discrimination that occurs at the exo-site, help distinguish between a G and oxoG in Fpg/MutM.

Acknowledgments

The material presented in this chapter was written by Arthur J. Campbell with suggestions and revisions from Professor Carlos Simmerling and Amber Carr. Nikita A. Kuznetsov, Grigory V. Mechetin, Vladimir V. Koval, Olga S. Fedorova and Dmitry O. Zharkov conducted the experimental studies. Arthur J. Campbell, Christina Bergonzo, Lin Fu, Haoquan Li and Carlos Simmerling performed the simulations and data analysis.

5.1 *Introduction*

Elaborate cellular repair processes have evolved to combat various types of DNA damage, one of which is the base excision repair pathway (BER).[1] In human cells, BER glycosylases usually process lesions that form between 100 to 10,000 lesions/day.[179,180] The repair processes of BER are performed with high fidelity and little or no energetic consumption. Fewer than a dozen human BER glycosylases have been determined that work to excise these lesions daily.[100] These glycosylases along with their bacterial counterparts share similar structural features: an intercalating aromatic/aliphatic wedge that enters through the minor groove, an overall bending/kinking of the DNA resulting from binding, and a base eversion process that flips the damaged nucleobase out of the duplex and into an active site for excision.[181-186] One of these glycosylases is human 8-hydroxyguanine DNA-glycosylase (hOGG1). The functional bacterial analog of hOGG1 is formamidopyrimidine DNA-glycosylase (Fpg/MutM), which excises the 8-oxo-guanine (oxoG) lesioned nucleobase as well as Fapy-G, Fapy-A and 5-OH-C.[187,188] Fpg/MutM is an ideal model system to study glycosylases in general, as a wide range of experimental and computational information is available for this system (Figure 5-1).[66,70,139,189,190]

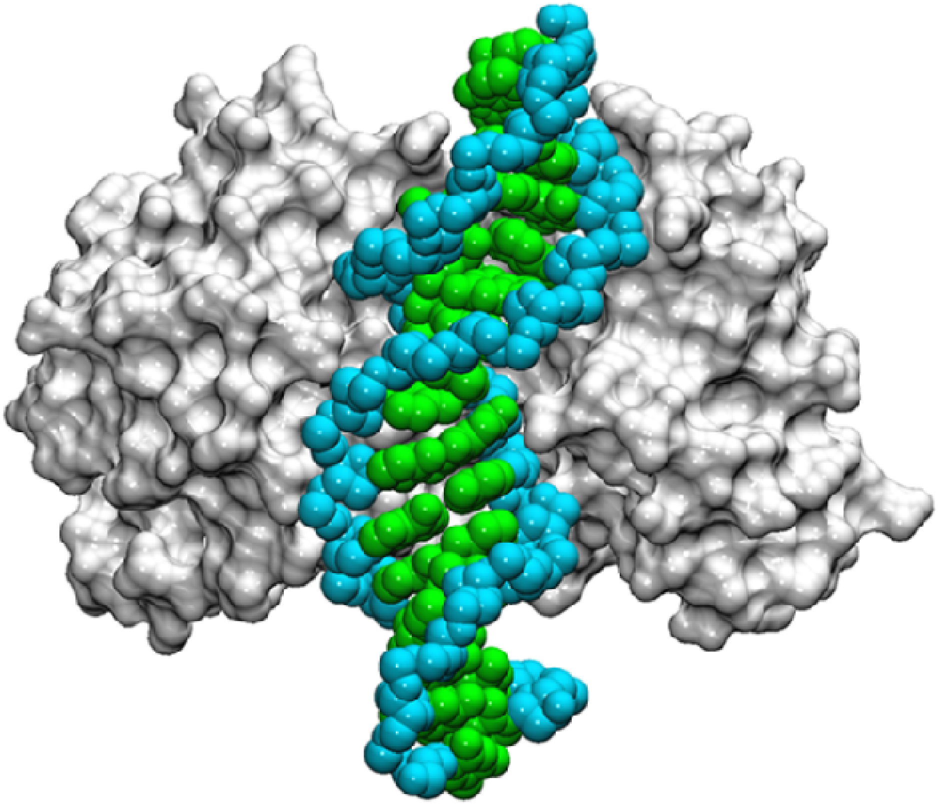


Figure 5-1. Equilibrated Fpg/DNA structure in which 8-oxo-G is intrahelical. Fpg is colored silver, the DNA backbone is blue, and the DNA bases are green.

Oxidative DNA damage is one of the most common types of damage known to affect the genome.[4,191] The main source of oxidative DNA damage is from free radicals of oxygen. It is paradoxical that oxygen, an essential element needed for cellular survival, could potentially be so harmful. Normal cellular activities, such as aerobic metabolism, can form reactive oxygen species (ROS), as can external factors, such as exposure to UV radiation.[192] One common product of ROS of particular interest is oxoG, which is formed from the oxidation of a normal guanine base, as shown in Figure 5-2.[193]

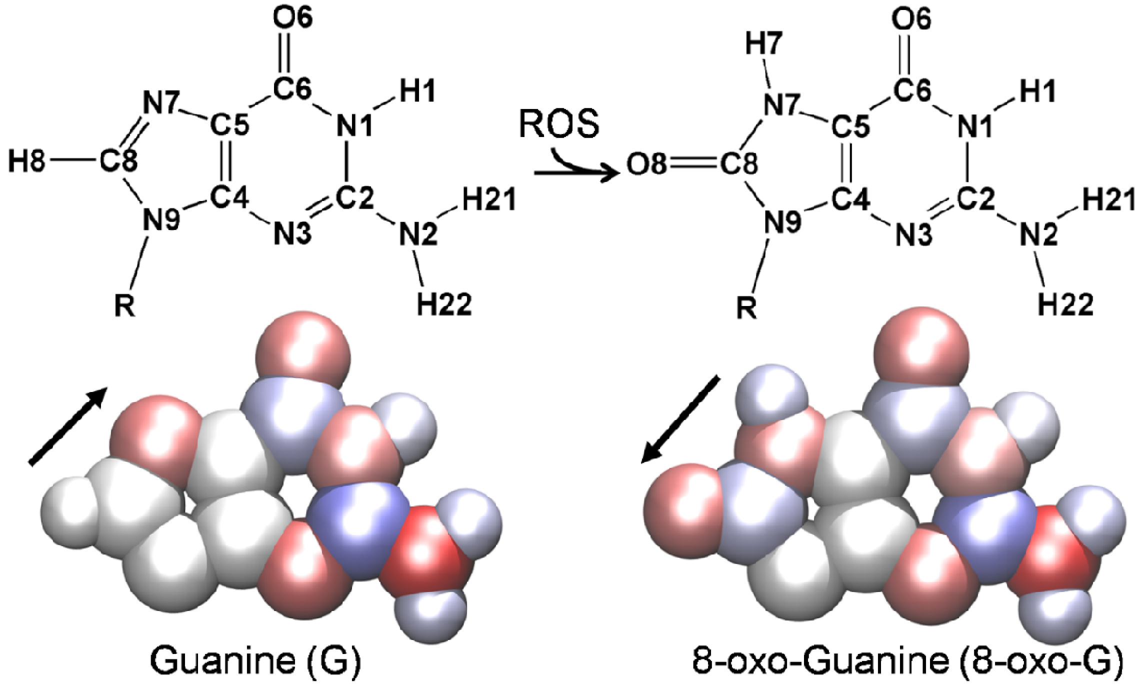


Figure 5-2. Atom names of Guanine (top left) and 8-oxo-Guanine (top right). The chemical differences between the two nucleobases can be compared by their electrostatic differences; bottom left (G) and bottom right (8-oxo-G). The colors for the electrostatics representation are red, white and blue for electronegative, neutral, and electropositive, respectively. Deeper variations of red and blue signify a greater extent of electronegativity or electropositivity. Arrows on the electrostatic potential figures indicate the local dipole direction (pointing towards the negative direction) where the damaged site is located.

Eukaryotic cells contain between 0.07 and 145.25 oxoG lesions per 10^6 nucleotides, which must be efficiently located and excised to maintain genomic integrity.[194] If the oxoG lesioned nucleobase remains uncorrected, it may lead to a G:C to T:A transversion mutation, which have been shown to be associated with cancer.[4,5]

The oxoG lesion is structurally similar to a normal guanine (G) nucleobase (Figure 5-2). The main difference is in two atoms on the major groove face in B-form duplex DNA. Whereas oxoG has an O8 and an H7 on the face, G has an H8 and a lone pair on N7. This difference of two atoms alters the electrostatics of oxoG compared to G (Figure 5-2, bottom). These electrostatic differences on the major groove face result in oxoG preferentially favoring the syn conformation resulting in the formation of specific hydrogen bonds with an adenine nucleobase in a Hoogsteen base pair manner upon a single round of replication, thus making a transversion mutation possible, (Figure 5-3).[195]

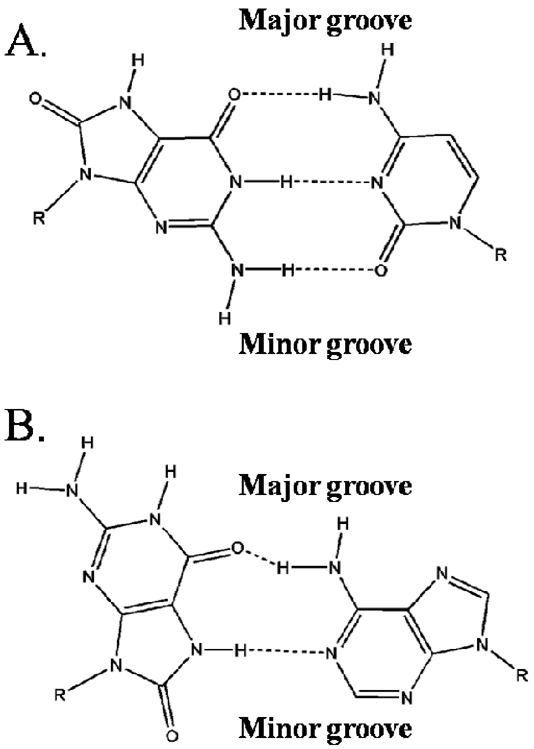


Figure 5-3. A.) Watson-Crick hydrogen bond formation between an oxoG and cytosine where both nucleobases are in the anti glycosidic conformation B.) Hoogsteen base pair formation between an oxoG and an adenine. oxoG and adenine are in a syn and anit glycosidic conformation respectively.

The BER DNA glycosylases that target oxoG as one of their substrates in eukaryotic and bacterial cells are hOGG1 and Fpg/MutM respectively. These glycosylases are functional analogs that only excise the oxoG lesion when opposite a cytosine in duplex DNA (Figure 5-3A). Although hOGG1 and Fpg/MutM have different tertiary structures, they contain analogous features that perform similar functions. Both enzymes have aromatic intercalating wedges (Tyr203 and Phe113 for hOGG1 and Fpg/MutM, respectively) that enter through the minor groove at the site of damage.[71,184] They also share a feature by which specific arginines (154/204 for hOGG1 and 112 for Fpg/MutM) enter through the minor groove and recognize the hydrogen bond acceptor atoms N3 and O2 of the orphan cytosine opposite oxoG.[184,189] The hydrogen bonds that form between the arginine residue(s) of each enzyme and the orphan cytosine is specific only for a cytosine and would not be accommodated by any other nucleobase.

In addition to similar probing mechanisms, hOGG1 and Fpg/MutM also share common methods for removing damage. Upon binding DNA, both enzymes force the damaged nucleobase to evert out of the duplex and into an active site for excision. The catalytic mechanisms of oxoG excision by hOGG1 and Fpg/MutM have been well characterized.[67,68,184,196-198] Both enzymes carry out nucleophilic attack on the C1' atom of the deoxyribose sugar of oxoG followed by β and δ elimination.[68,184] A crystallographic structure of the catalytically inactive E2Q mutant of Fpg/MutM with oxoG trapped in the active site loop just before excision was first shown by the Verdine group.[20] In this structure, oxoG is held in the active site by forming several hydrogen bonds between the active site loop and the O6 atom of the extra-helical nucleobase.[70]

As mentioned, both enzymes have intercalating wedges and accompanying residues that make specific hydrogen bonds to the orphan cytosine which come through the minor groove. In the 2F5O crystallographic structure, Fpg/MutM is complexed to DNA and interrogating an intra-helical G opposite its Watson Crick partner C; no Fpg/MutM residues contact the major groove face of G. This structure suggests that if an oxoG were intra-helical and opposite C in a Watson Crick hydrogen bonding fashion, no Fpg/MutM residues would contact the damaged side therefore Fpg must recognize the damage by inducing some change in the enzyme, DNA or both. This hypothesis was confirmed by the recent crystallographic structures of Fpg/MutM in which the active site loop was either mutated or deleted which destabilizes the extra-helical state and stabilizes the intra-helical. These intra-helical cyrstallographic structures confirm that an indirect readout mechanism must be employed by this enzyme to recognize damage.[184,189]

Two recent articles have been published on Fpg/MutM suggesting that eversion proceeds through the minor groove pathway.[72,73] Although both works were highly insightful they did not explore other possible routes. One work revealed a crystallographic structure of Fpg/MutM complexed to DNA with a partially everted G in the minor groove. The enzyme in this structure was covalently bound to the DNA using crosslinking technology where a N173C mutation was made and a crosslink was formed between residue 173 and the DNA backbone.[73] This work assumes that N173 does not play a significant role in base eversion nor does it take into account that an N173C crosslink may sterically block the major groove pathway. The other work gives energetic profiles for eversion through the minor groove with an overall barrier of ~13 kcal/mol for everting oxoG.[72] This is a high barrier to overcome for an enzyme which only uses thermal energy to facilitate eversion.[72] Therefore, more rigorous work is needed to investigate in which direction eversion occurs for Fpg/MutM, and to elucidate the mechanism by which discrimination occurs.

It has been hypothesized that hOGG1 everts oxoG out of the duplex through a major groove pathway and has a stable exo-site that is thought to distinguish between a G and an oxoG, thus allowing oxoG to continue everting while prohibiting G from further extrusion.[165] We hypothesize that Fpg/MutM shares a similar mechanism to hOGG1 where base eversion occurs through the major groove and a stable exo-site is used to discriminate between the two nucleobases. In this work, Fpg's discriminatory behavior between G and oxoG will be discussed through the use of experimental and computational studies.

5.2 *Methods*

5.2.1 Initial structure generation for 8-oxo-G.wt and G.wt

The Amber 9 and 10 suites of programs were used for all calculations in this work.[27] For all systems the ff99SB force field was used with the parmbsc0 DNA backbone parameters.[78,81] The alpha and gamma backbone parameters were altered to include the parmbs0 parameters for DNA.[81] The parameters used in this work for oxoG were based on the parameters developed by Miller et al.[164] Four initial structures were generated based on two systems: wild type Fpg bound to DNA interrogating oxoG (oxoG.wt) and wild type Fpg bound to DNA interrogating G (G.wt).

Initial structures for the intra-helical and extra-helical states were generated from two crystallographic structures, pdb IDs 2F5O and 1R2Y (Fpg/MutM interrogating an intra-helical and extra-helical nucleobase respectively).[70,71] In the 2F5O crystallographic structure, Fpg/MutM is bound to DNA and is interrogating an intra-helical Watson Crick G:C base pair. In the 1R2Y crystallographic structure, Fpg/MutM is bound to DNA which contains an oxoG that is fully extra-helical and in the active site loop. Since there are differences between the two crystallographic structures (described further below) both structures were modified to be equivalent in their primary structure. The DNA sequence of 2F5O was used as the standard (Table 5-7). Any disordered nucleobases in the 2F5O sequence were built in with MoilView using the coordinates from a standard duplex made in Nucgen.[27] Any nucleobase that was not identical in 1R2Y to the sequence in Table 5-7 was deleted, keeping the sugar, phosphate and part of the base. The correct atoms for Table 5-7 were then built using Amber's tleap module. Any missing nucleobases in the 1R2Y structure were modeled in using MoilView based on a standard duplex made in Nucgen. The 2F5O crystallographic structure does not have an ordered active site loop, using MoilView this was modeled using the coordinates from the ordered active site loop from the 1R2Y crystal structure.[70,71] The E3Q inactivating mutation in 1R2Y was reversed to the wildtype E3 sequence. Although the experimental conditions included protein-DNA crosslinks, these were not included in the published coordinates. These simulations correspond to the un-crosslinked species. The back mutations where G was mutated to an oxoG and oxoG was mutated to a G was executed by deleting the orginal nucleobases H8 or O8/H7 atoms for G and oxoG respectively and then using Ambers tleap module to build in the mutation.[27]

Table 5-7. DNA sequence and residue numbering of all systems used in this work.

274	**275**	**276**	**277**	**278**	**279**	**280**	**281**	**282**	**283**	**284**	**285**	**286**	**287**	**288**	**289**	
A 5'	**G**	**G**	**T**	**A**	**G**	**A**	**T**	**C**	**C**	**G**	**G**	**A**	**C**	**G**	**C 3'**	
	C 3'	**C**	**A**	**T**	**C**	**T**	**A**	**G**	**8O(G)**	**C**	**C**	**T**	**G**	**C**	**G**	**T 5'**
	305	**304**	**303**	**302**	**301**	**300**	**299**	**298**	**297**	**296**	**295**	**294**	**293**	**292**	**291**	**290**

5.2.2 Structure equilibration

After both intra- and extra-helical structures were mutated to chemically equivalent systems for oxoG.wt and G.wt, they were minimized, equilibrated and relaxed further in unrestrained dynamics.

Two sequential rounds of minimization were run for 200 steps of steepest descent where all heavy atoms of the systems were held in place with a restraint force constant of 5.0 kcal/mol-Å^2. After minimization the systems were then solvated with ~10,000 TIP3P explicit solvent water molecules in a truncated octahedron periodic box with an 8Å buffer. After the systems were solvated they were further minimized over five consecutive 1,000 step rounds with the steepest descent method where the restraint force constant was incrementally relaxed from 10.0, 5.0, 2.5, 1.0 and 0.0 kcal/mol-Å^2 respectively on all heavy atoms of the DNA and Fpg.

Five consecutive rounds of equilibration were employed after minimization under constant pressure at 1 atm with a timestep of 2fs where all bonds involving hydrogen were constrained using SHAKE.[121] Temperature was controlled by using the weak-coupling algorithm.[122] In the first round of equilibration all atoms of the DNA and enzyme were frozen while all water molecules were allowed to relax for 50ps. Also during the first step of equilibration the systems were slowly heated from 0 to 300K. In the second, third and fourth steps of equilibration the temperature was held constant at 300K. In the second step of equilibration all atoms on the DNA and Fpg/MutM were restrained with a restraint force of 5.0 kcal/mol-Å^2 for 50ps. In the third step the restraints on the DNA and Fpg/MutM were decreased to 1.0 kcal/mol-Å^2 for 50ps and in the fourth step the restraints were completely removed and the entire systems was relaxed in unrestrained dynamics for 50ps. In the fifth and last step the systems was slowly heated from 300K to 330K over the course of 500ps since Fpg/MutM is a thermophile and 330K is the biologically relevant temperature.[20] After the last step of equilibration the systems were simulated for ~20ns in unrestrained dynamics. The systems were coupled to a thermostat and barostat at 330K and 1atm, respectively.[122] After the intra-helical and extra-helical structures were equilibrated the final structures were used as starting structures for the remainder of the simulations.

5.2.3 Running partial nudged elastic band (PNEB)

To develop a low potential energy pathway between the intra-helical and the extra-helical states for Fpg/MutM bound to DNA when interrogating a G or oxoG, nudged elastic band method was employed.

To sample major groove eversion pathway, midpoint beads were used to guide the everting base in the proper direction. The initial beads for oxoG.wt and G.wt were generated from the endpoint structures discussed in the previous section, as well as midpoint structures which were partially everted to guide the system through the major groove. Partially everted structures were required to seed the eversion path of G.wt and oxoG.wt through the major groove since without seeding the path NEB would take the most direct path (the minor groove direction) regardless if it is energetically less favorable. To develop the midpoint beads for G.wt and oxoG.wt, a short restrained molecular dynamics simulation was run where G and oxoG were separately forced to evert out through the major groove from the intra-helical position by applying an harmonic restraint set to an eversion angle of ~195 degrees. This simulation was run for 30ps with a timestep of 2fs. The temperature was maintained at 330K using the weak coupling algorithm. To avoid creating any distortion in the enzyme, the backbone atoms of the Fpg/MutM were restrained with a .10 kcal/mol-Å^2 force constant. To avoid circumvent any artificial distortion in the DNA caused by everting a single nucleotide out of the duplex, the entire strand opposite the everting base strand was restrained along with residues 293-297 (Table

5-7) with a force constant of .10 kcal/mol-Å^2. On the restrained DNA residues, all heavy atoms of the sugar and the phosphorus atom were restrained. Using a force constant of .0305 kcal/mol-degrees^2 on the eversion angle was found to create the least amount of artificial distortion when everting the G and oxoG nucleobases (Table 5-7, residue 297) out of the duplex and in the positive Phi direction (out through the major groove).

To seed the NEB pathway for oxoG through the major groove 32 beads were found to give enough resolution of the eversino pathway for structural anlayis. The last structure from the short restrained molecular dynamics simulation was used as the central 30 beads to keep the oxoG in the major groove pathway. The intra-helical and extra-helical structures generated from the previous structure equilibration section were used as the endpoints beads for the oxoG major groove direction. For the G major groove pathway, 31 beads were needed to obtain a major groove pathway in NEB that gave the lowest possible potential energy path. A structure from the short restrained molecular dynamics simulation was chosen where the Watson Crick hydrogen bonds were broken and this was used as a central bead along with 15 intra-helical and 15 extra-helical endpoint bead structures generated in the equilibration section above, comprising of the 31 beads in total for the G.wt system that were initially used for NEB. The total number of beads chosen for G.wt was found to give enough resolution of the major groove pathway for structural analysis.

Since the minor groove pathway was the most direct of the two pathways (minor/major) no seeding was required. To sample the minor groove NEB pathway for oxoG.wt and G.wt, 9 intra-helical and 19 extra-helical beads were used which were from the previous structure equilibration section comprising of 38 beads in total for each minor groove system. The total number of beads used in the minor groove pathway was found to give enough resolution for structural analysis.

The protocol for simulated annealing used in this work was adopted from Mathews & Case protocol.[163] All bonds involving hydrogen were constrained using SHAKE and volume was held constant.[121] A Langevin thermostat and a nonbonded cutoff of 8.0 Å was used in all six steps of annealing.[199] Six steps of heating, cooling and equilibration were used. In NEB, the initial path must be optimized to find the minimum energy pathway. Therefore, six consecutive rounds of annealing were performed on the 31/32 (major groove) or 38 (minor groove) initial beads described above. Springs were applied between neighboring beads to keep them evenly spaced along the pathway of interest. A decoupling by a tangent based on neighbor bead positions or energies removed the effect of the springs during the optimization of each bead by the force field. The NEB code in Amber's multisander functionality allowed the NEB forces to be applied to a subset of the system's atoms.[27,91] For all systems discussed here, NEB forces were applied to all heavy atoms of the protein and DNA.

In the first round of annealing for NEB, all 31/32 (major groove) or 38 (minor groove) beads were simulated for 80ps with a timestep of 2fs. The spring constant was set to 10 kcal/mol-Å^2. The system temperature was held at 330K with a collision frequency of 100.0 ps^{-1}.

The spring constant was then increased from 10 to 50 kcal/mol-Å^2 in the second round of NEB annealing to force the beads to sample the pathway of interest. Each bead was simulated for 500ps with a timestep of 1fs at 330K with a Langevin collision frequency of 75.0 ps^{-1} which was held constant for the remaining rounds of annealing.[199]

Next the spring constant was decreased from 50 to 25 kcal/mol-Å^2 for the third round of annealing and the temperature was increased to lower the energetic barriers to sample multiple conformations. In this stage the system was incrementally heated from 330K to 386K over the course of 60ps with a timestep of 2fs.

In the next stage of NEB the temperature of the system was held constant at 386K for 220ps with a timestep of 1fs. The spring constant in this round was held at 25 kcal/mol-Å^2 to allow the system to sample multiple conformations.

In the fifth round of annealing the spring constant was still kept at 25 kcal/mol-Å^2 while the system was incrementally cooled from 386K to 330K over the course of 100ps with a timestep of 2fs to bring the system temperature back to the biologically relevant state.

In the final round of NEB, the system was equilibrated, allowing the system to relax, at 330K over the course of 500ps with a timestep of 2fs and a spring constant at 25 kcal/mol-Å^2.

5.2.4 Umbrella Sampling

Umbrella sampling was employed to obtain free energy profiles of eversion. After running the final round of NEB, which developed low potential energy paths going from the intra-helical to the extra-helical position, the last stage of annealing for the four systems: oxoG.wt (major/minor) and G.wt (major/minor) were sorted based on two reaction coordinates (Figure 5-4, top left), using a measure of eversion and the glycosidic torsion angle (Figure 5-5). The starting structures for each of the four systems for umbrella sampling were taken from this last stage of NEB, by extracting the closest structures based on a predefined grid spacing of 10 degrees with a limit of 10 degrees from the structure to the closet grid point (Figure 5-4, top right) (296, 349, 201 and 218 total windows were used for oxoG.wt major groove, G.wt major groove, oxoG.wt minor groove and G.wt minor groove respectively) which represents each starting structure for each separate umbrella sampling window. Each window was restrained to its respective grid point coordinates with a restraint force constant of .0609 kcal/mol-degrees^2 which was found to give the ideal amount of overlap based on the predefined grid points which is essential for 2D-WHAM, (Figure 5-4, bottom right).

The G.wt (major/minor) and oxoG.wt (major/minor) systems were each simulated for 500ps with a timestep of 2fs. The temperature of each window was held constant at 330K and volume was held constant. An 8Å nonbonded cutoff was applied to all simulations. All bonds involving hydrogen were constrained using SHAKE.[121]

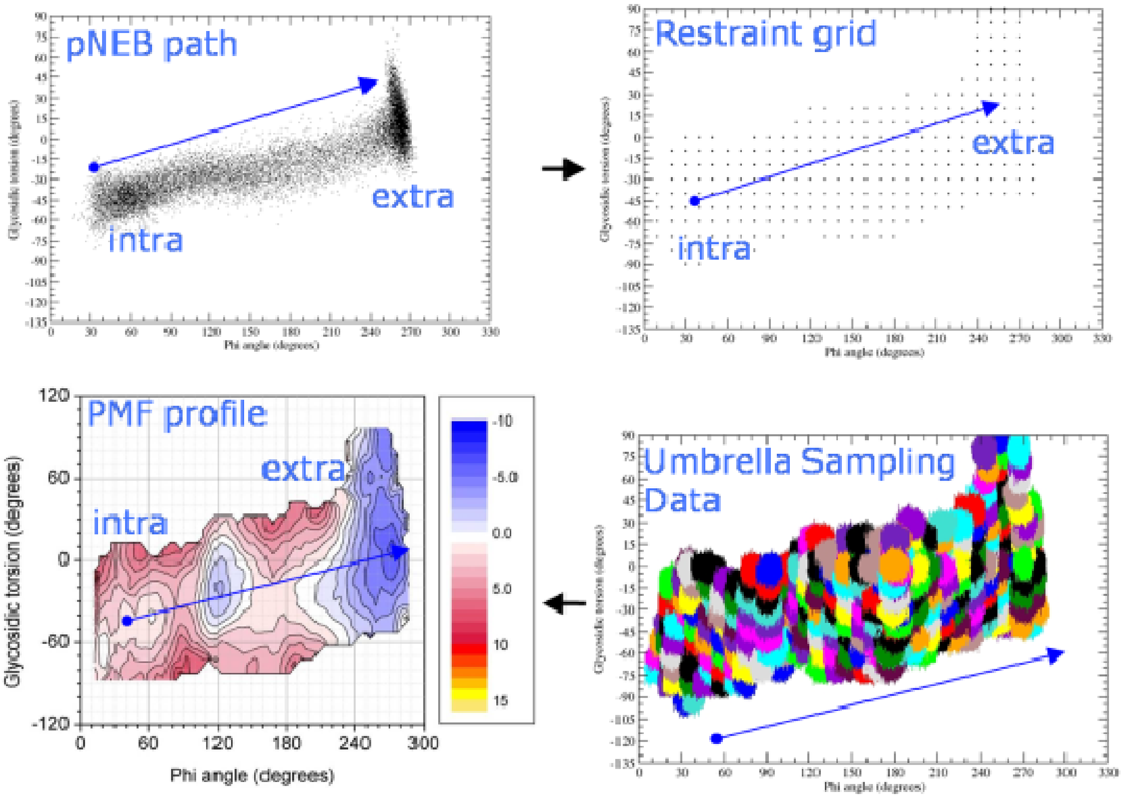


Figure 5-4. Workflow diagram for 8-oxo-G.wt through the major groove. All four figures are based on the two reaction coordinates eversion and glycosidic torsion angle (x-axis and y-axis respectively). Last equilibration step of PNEB data (top left). Structures were extracted from the final step of PNEB data (top left) based on predefined grid points (top right). Each grid point (top

right) represents the starting structure for our umbrella sampling simulations (bottom right). From the umbrella sampling simulations we were able to extract the two dimensional PMF profiles using 2D-WHAM (bottom left).

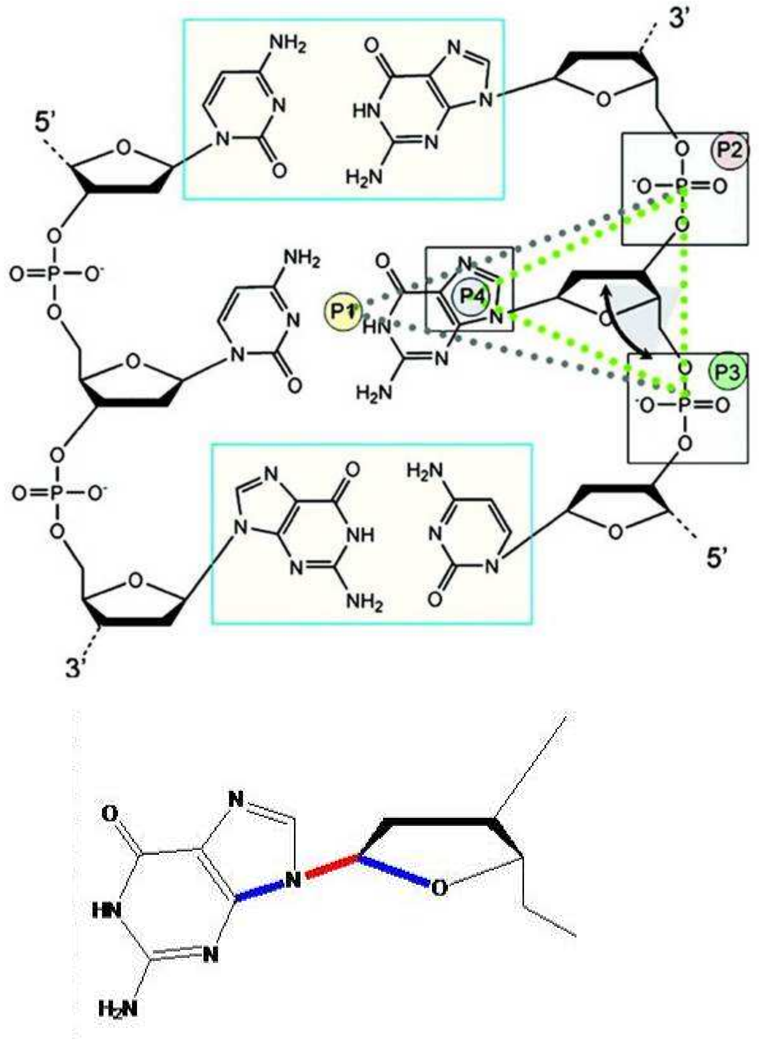


Figure 5-5. Angle of base eversion, top. Glycosidic torsion angle, bottom The eversion angle is separated into four points: point 1 is the center of mass (COM) of the heavy atoms of the two flanking base pairs to the everting nucleobase, point 2 is the COM of the 3' phosphate to the everting nucleobase, point 3 is the COM of the 5' phosphate group to the everting nucleobase and point 4 is the COM of the heavy atoms of the 5 member ring of the everting nucleobase.

5.2.5 Generating and simulating the F113A mutation

The F113A mutation was made for the G.wt and oxoG.wt major groove systems, (named G.F113A and oxoG.F113A respectively). To sample the identical major groove pathways for the G.wt and oxoG.wt systems sampled from NEB, the mutations were based on all initial starting structures that were extracted from NEB as initial structures for the G.wt and oxoG.wt major groove pathways respectively. To execute these mutations, all atoms of the sidechain of F113 were removed and the correct atoms for alanine were built in using Amber's tleap module.[27] The same protocol was then used that is outlined above for oxoG.wt and G.wt umbrella sampling to obtain PMF profiles for G.F113A and oxoG.F113A.

Each window for oxoG.F113A and G.F113A was equilibrated for 100 ps with a timestep of 2 fs to allow the mutated system to relax. All heavy atoms of the DNA and Fpg were restrained with a restraint force constant of 10 kcal/mol and the explicit solvent was allowed to move since we removed the bulky aromatic wedge leaving a cavity which water may enter. In this stage the temperature was held constant at 330K using a Langevan thermostat. After equilibration the same protocol was used that is outlined above for oxoG.wt and G.wt umbrella sampling to obtain PMF profiles.

5.2.6 Removing the NH2 charges on N173

We were interested in measuring the effects of removing the N173 interactions in the major groove pathway. To accomplish this the charges on the amino group of N173 for the oxoG.wt and G.wt major groove systems were removed, (named oxoG.N173' and G.N173' respectively). Since we were interested in sampling the identical major groove pathways for the oxoG.wt and G.wt systems sampled from NEB, the mutations were based on all ~300 initial starting structures that were extracted from NEB as initial structures for the oxoG.wt and G.wt major groove pathways respectively. To execute these mutations, the charges on ND2, HD21 and HD22 atoms for oxoG.N173' and G.N173' were set to zero in their respective topologies. The same protocol was then used that is outlined above for oxoG.wt and G.wt umbrella sampling to obtain PMF profiles for oxoG.N173' and G.N173'.

5.2.7 Generating PMF profiles

The eversion and glycosidic angles were written for each timestep of each Umbrella Sampling simulation, (Figure 5-4, bottom right). After the umbrella sampling simulations were complete the two dimensional Potential of Mean Force (PMF) profiles were extracted from the eversion and glycosidic data written using two Dimensional Weighted Histogram Analysis (2D-WHAM) analysis (Figure 5-4, bottom left) (freely available at http://membrane.urmc.rochester.edu/).

To calculate the PMF profiles for oxoG.wt (major/minor pathways), G.wt (major/minor pathways), oxoG.N173' and G.N173', 2D-WHAM was employed. The input for 2D-WHAM in this work was the eversion and glycosidic torsion angles written at each time step for each umbrella sampling window along with the respective restraint force constant (.0609 kcal/mol-degrees^2) holding each window to the assigned grid point and the temperature the umbrella sampling calculations were simulated at (330K).

All PMF profiles were all generated using Origin 8.0 edition where the intra-helical position was set to zero to compare the relative stability differences among the systems as eversion proceeds.

5.2.8 oxoG.F113W mutation details

Several of our FRET experiments included the F113W mutation. To validate that the tryptophan is a reasonable intercalating analog, we modeled in the F113W mutation in the intra-helical oxoG.wt system. After the last step of equilibration of the intra-helical oxoG bound to Fpg we mutated F113W. The backbone and part of the F113 (atoms HD2, CD2, CE2, HE2, CZ and CD1) were kept to maintain the same plane as the wild type F113 in the mutation. The sidechain atoms that were kept were then renamed (HD1, CD1, NE1, HE1, CE2 and CD2) for the mutant. The F113W mutation was then performed using Amber's tleap module.[27] After the mutation we minimized the structure for 1,000 steps of steepest descent followed by 20ns of unrestrained dynamics under the same conditions outlined above for oxoG.wt.

5.2.9 Energy decomposition

To measure the residue to residue interactions among the simulations the nonbonded interactions were separately calculated. The nonbonded interactions calculated consists of the electrostatic and van deer Waals interaction terms.

5.2.10 Binning

The raw data for all contour figures (except the PMF profiles) were plotted in the same manner. For each system, the eversion angle, glycosidic torsion angle and variable data point was analyzed for each picosecond from their respective umbrella sampling simulations. The average variable data point was calculated for every 5 by 5 degree bin in the measurement. All measurements were then plotted in Origin 8.0.

5.3 *Results and discussions*

5.3.1 Energetic barriers to eversion of 8-oxo-G and G in Fpg

In this work we are interested in computationally and experimentally comparing the eversion process of an undamaged guanine (G) nucleobase to the eversion of a damaged oxoG nucleobase in duplex DNA when bound by Fpg, (Figure 5-1 and Figure 5-2). As mentioned, Fpg/MutM may share a similar mechanism to eversion as hOGG1 where base eversion is thought to occur through the major groove and discrimination occurs extra-helically. To test this hypothesis, first the free energy of the transition of the nucleobase from the intra-helical to extra-helical state is needed then any interactions that occur during this process can be measured to see what is contributing to any energetic differences. Previous crystallographic work has been able to provide snapshots along the eversion process.[70-72,189] Of particular interest are two crystallographic structures of Fpg/MutM bound to DNA that define two states: one state in which the nucleobase is intra-helical (PDB 2F5O)[71] and another in which it is extra-helical (PDB 1R2Y)[70]. Although these two structures are useful for understanding two biologically relevant states, they give an incomplete picture of how Fpg discriminates against G and for an oxoG during eversion since they are only two snapshots along a dynamic process where discrimination must occur. This work will shed light on what happens in the low-populated higher-energy states which occur between these two snapshots when everting a G and an oxoG when complexed to Fpg/MutM to understand how discrimination occurs. To accomplish this nudged elastic band method coupled with umbrella sampling will be used to explore the eversion pathways for both nucleobases to understand Fpg/MutMs mechanism for discrimination.

Uncatalyzed base eversion in naked DNA has been experimentally found to occur on the millisecond timescale[140], which is currently outside the computational sampling capabilities of unrestrained Molecular Dynamics (MD) therefore enhanced sampling techniques are needed to study this process. Previous work investigating base eversion in Fpg/MutM have used steered molecular dynamics (SMD) along with targeted molecular dynamics (TMD) to investigate base eversion through the minor groove.[72]

Since our interests lie in understanding the energetics of the system during eversion and the interactions that occur in the transition from the intra-helical to extra-helical positions for both the minor and major groove directions, one computational technique that can be used is the nudged elastic band (NEB) method, coupled with two-dimensional umbrella sampling (US) (see Methods section).[91,94-96] NEB is a reaction-coordinate-independent method used to find the minimum potential energy path between two well-defined states of a system, in this case the intra-helical (PDB 2F5O)[71] and extra-helical (PDB 1R2Y)[70] positions of Fpg bound to duplex DNA when separately probing a G or oxoG.

One advantage to using NEB as opposed to SMD or TMD is that the initial starting structures can be seeded to investigate multiple pathways of interest such as the minor and major groove pathways. Umbrella sampling can then be used to obtain the free energy of the low potential energy path generated in NEB (see Methods section).

In this work the US is done two-dimensionally using an eversion angle and the glycosidic torsion angle of the everting nucleobase as the two reaction coordinates along which the free energy path of the system is determined, (Figure 5-5). These two reaction coordinates were chosen because the eversion angle was previously shown in our laboratory to be a reliable measure of eversion capable of measuring through both the minor or major grooves[200], and the glycosidic torsion was chosen because the intra-helical nucleobase is in the syn position (PDB 2F5O)[71] and the extra-helical nucleobase is in the anti position (PDB 1R2Y), therefore a transition must occur somewhere along the eversion path. By using the two reaction coordinates chosen for US, specific energy minima and maxima can be linked to specific structural interactions that occur during eversion.

Since we were interested in testing our hypothesis that the mechanism for base eversion through the major groove is conserved between hOGG1 and Fpg/MutM, the energetic barriers to eversion of G and oxoG in duplex DNA when complexed to Fpg/MutM were measured through the major and minor grooves to understand in which direction eversion occurs.(Figure 5-6 and Figure 5-7) The minor groove PMF's (having an overall barrier to eversion of ~8 and ~11 kcal/mol for G and oxoG respectively) were found to be consistent with previous work by the Verdine/Karplus groups (having an overall barrier to eversion of ~12 and ~13 kcal/mol for G and oxoG respectively). The major groove pathway was found to be the lower energetic pathway (~3-4 kcal/mol) for the eversion of oxoG, which is consistent with the mechanism of the enzyme, which does not consume energy to accomplish eversion or excision.[72] When sampled through the major groove path, the energetic barrier to eversion for G and oxoG was found to be ~7 and ~3-4kcal/mol repectively, which is significantly lower than the minor groove path for oxoG and consistent with an enzyme that relies solely on thermal motions and binding energy to facilitate eversion of oxoG.[201] Since the major groove pathway in our work was found to be more favorable for both nucleobases, this pathway will be the focus of this work.

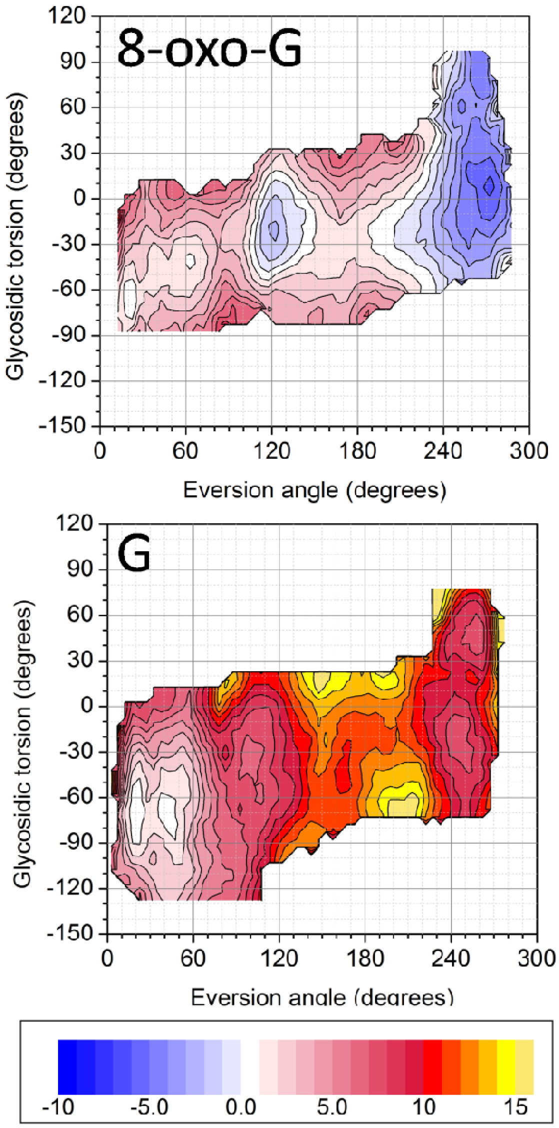


Figure 5-6. PMF profiles of 8-oxo-G (top) and G (bottom). The intrahelical region is represented by an eversion angle of less than ~20 degrees and a glycosidic torsion angle of ~-60 degrees. The extrahelical region is represented by a Phi angle of ~250 degrees and a glycosidic angle above ~-50 degrees. The legend is in kcal/mol and each contour represents 1kcal/mol.

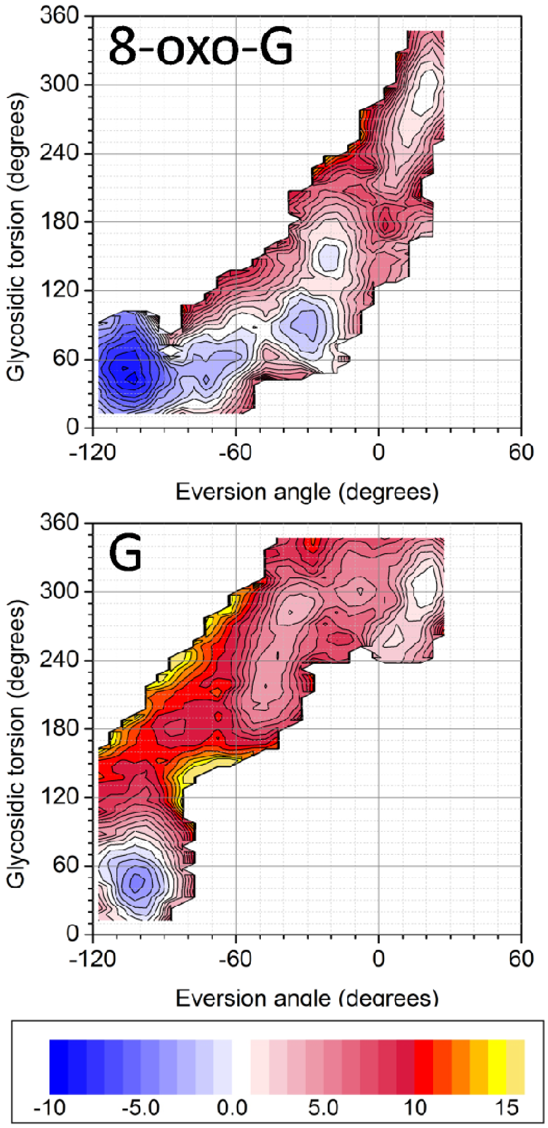


Figure 5-7. PMF profiles for 8-oxo-G (top) and G (bottom) through the minor groove when complexed to Fpg. The legend is in kcal/mol

The overall equilibrium for G and oxoG found in the energetic landscapes of both pathways is also consistent with the two crystallographic structures 2F5O[71] and 1R2Y[70], in which G favors the intra-helical state (2F5O) and oxoG favors the extra-helical state (1R2Y), (Figure 5-6). The equilibrium difference found for oxoG in Verdine/Karplus's computational work is not consistent with the crystallographic structure 1R2Y since the energy landscape in the computational work favors the intra-helical state and the crystallographic structure favors the extra-helical state in the active site loop.

When comparing the major groove energetic profiles of G and oxoG, the intra-helical and extra-helical states were found to be substantially different relative to one another. G favors the intra-helical position by ~-10kcal/mol with respect to its extra-helical state (Figure 5-6). For oxoG, the extra-helical position is favored by ~-7kcal/mol with respect to its intra-helical position. This is consistent with experimental data which indicates that excision only occurs for oxoG as compared to that of G.[70,71] Not only do the endpoints differ in the energetic landscapes in the major groove pathways but other regions differ as well which will be discussed in this work.

5.3.1 Base pair opening requires little energy for both nucleobases

To better understand the interactions that make the regions stable for each system, structures from the US simulations of both systems were inspected based on the two reaction coordinates, the eversion angle and the glycosidic torsion angle so they can be directly compared to their subsequent energy landscapes (Figure 5-5 and Figure 5-6).

We were interested in understanding where along the eversion process the Watson Crick (WC) hydrogen bonds break and if there is an energetic difference to breaking nondamaged vs. damaged basepairs. For the major groove eversion pathway of G.wt and oxoG.wt, the WC hydrogen bonds were found to break at an eversion angle of ~20 degrees while the G.wt hydrogen bonds break slightly later on the eversion pathway, Figure 5-8. Therefore the intra-helical region for the major groove pathway in this work is generally considered anything less than an eversion angle of 20 degrees.

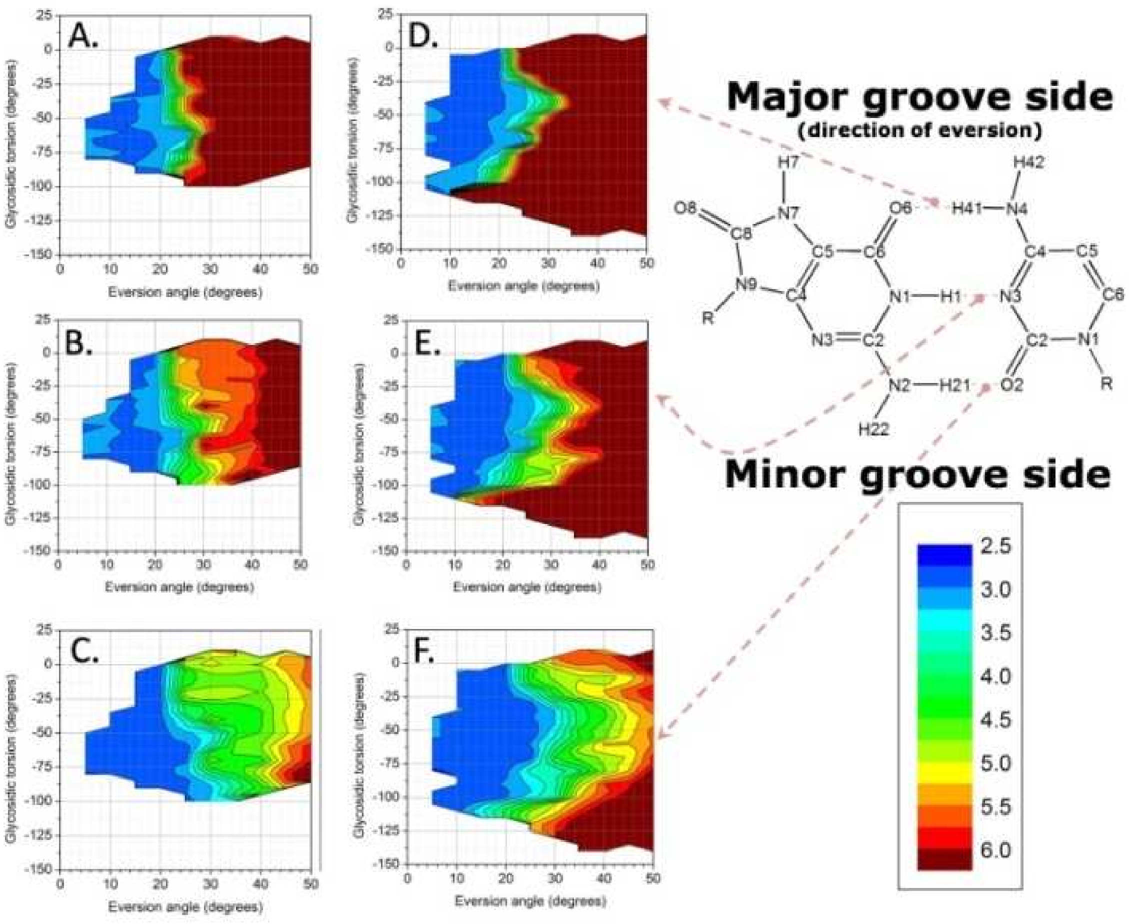


Figure 5-8. Distances corresponding to the three (oxo)G:C Watson Crick hydrogen bonds during eversion shown to an eversion angle of 50 degrees. The left column shows the oxoG WC measure and the right column shows the G WC measure. A/D corresponds to the O6 to N4 distance, B/E corresponds to the N1 to N3 distance and C/F corresponds to the N2 to O2 distance between (oxo)G and C respectively. Atom names correspond to top right figure. Legend is in Angstroms, each contour represents .25 Å.

To investigate if the slight differences in the WC hydrogen bond separation was a result of differences in the base pair buckling of G:C and oxoG:C, the buckling was measured using

CURVES analysis, Figure 5-9.[202] It was found that there is little difference in the extent of buckling in the intra-helical region of both systems, Figure 5-9.

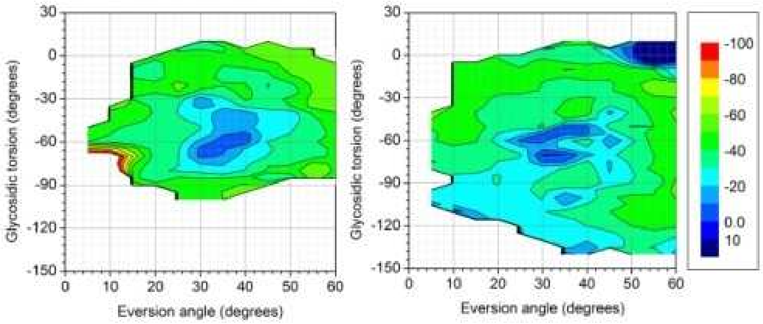


Figure 5-9. oxoG(G):C buckling during the initial stages of eversion (eversion angle up to 80° is shown) for oxoG.wt (left) and G.wt (right) Fpg. Significant increases in buckling are observed in the presence of the Phe wedge in the intra-helical region which is an eversion angle less than 20 degrees.

When comparing the WC base pair separation of both G.wt and oxoG.wt to their respective energy landscapes before and after an eversion angle of ~20 degrees, little or no energy is required to initially break the WC bonds for both systems, Figure 5-6 and Figure 5-8. This low energy barrier is a direct result of the buckling caused by Fpg/MutM's aromatic intercalating Phe 113 wedge. This was confirmed by mutating the aromatic wedge in both systems to an alanine (F113A), which resulted in a direct increase in energy upon the WC hydrogen bonds breaking suggesting that the local distortion caused from the wedge destabilizes both base pairs indiscriminately, Figure 5-10 (discussed further in chapter 4).

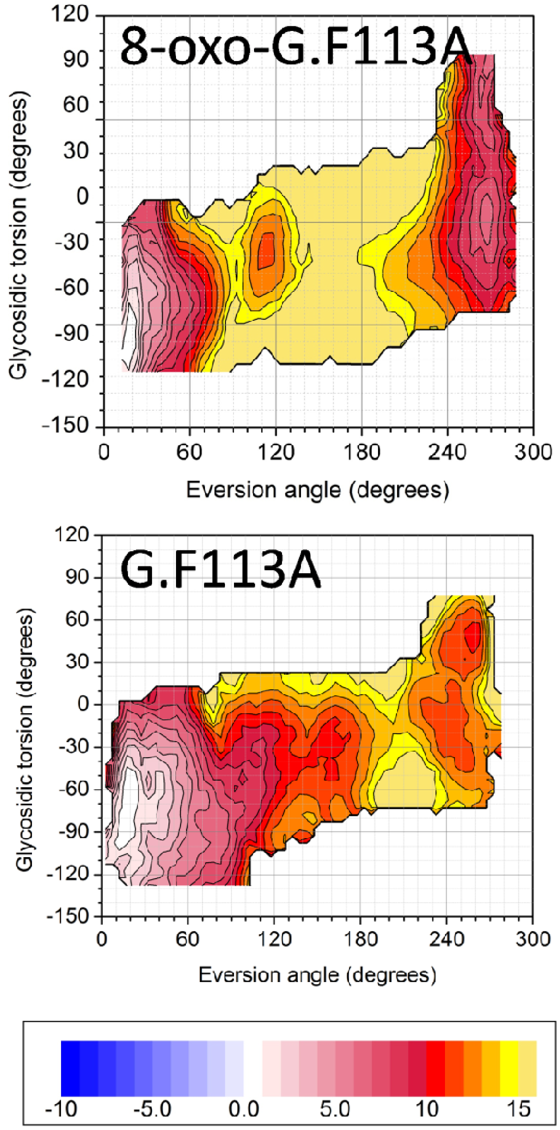


Figure 5-10. PMF profiles of oxoG.F113A (top) and G.F113A (bottom). The intra-helical region is represented by an eversion angle of less than ~20 degrees. The legend is in kcal/mol and each contour represents 1kcal/mol.

After both nucleobases leave the intra-helical region and begin the base eversion process in the major groove pathway, they first come into contact with Arg 263 on the Zinc finger of Fpg/MutM. When G is everting, Arg 263 forms a weak nonbonded interaction (electrostatic and van der Waals) of -1 to -2 kcal/mol until it reaches an eversion angle of ~80 degrees where it forms a more favorable cation-pi interaction, discussed further below (Figure 5-11). A result of the weak interaction between Arg 263 and G is that the Zinc finger can be drawn in closer to the DNA backbone to make a more favorable interaction with Lys 257 (Figure 5-12 and Figure 5-13).

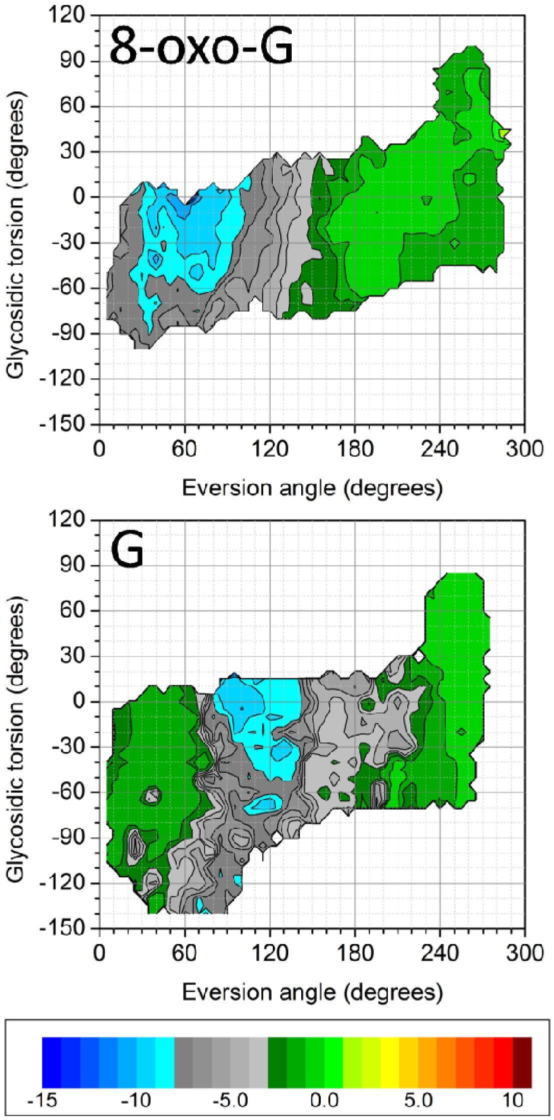


Figure 5-11. Interaction energy between the everting nucleobase (base only) and Arg263 during the eversion process. To confirm that there is a favorable interaction between Arg263 and the 8-oxo-G nucleobase we measured the distances between C8 of 8-oxo-G/G to the NH2 atom of Arg263, oxo-G (top) and G (bottom). The legend is in kcal/mol.

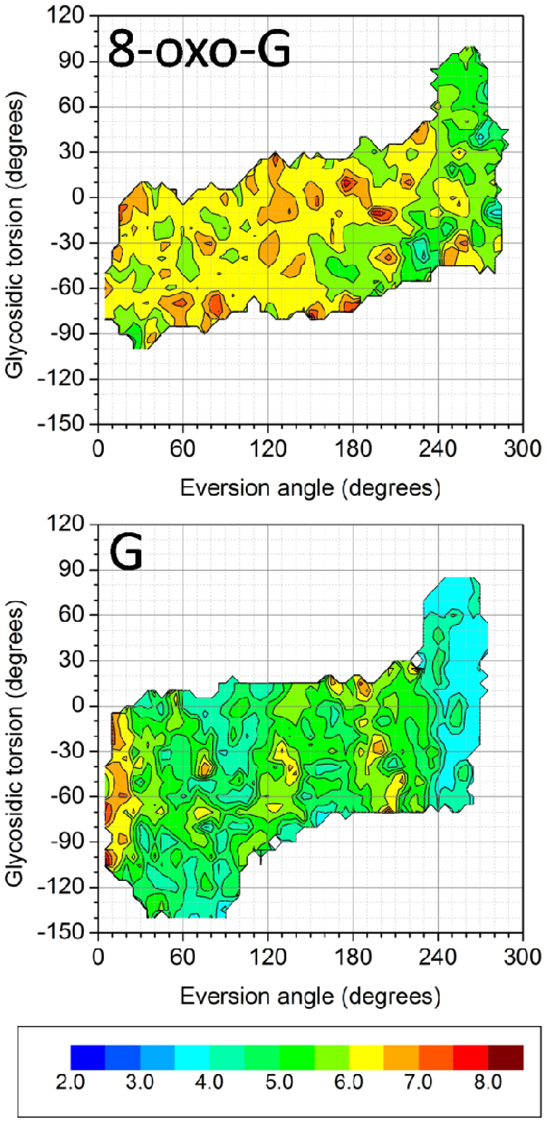


Figure 5-12. Distance measurement between the NZ of each Lysine 257 and the P atom of the second 5' phosphate for oxoG (top) and G (bottom). The legend is in Angstroms.

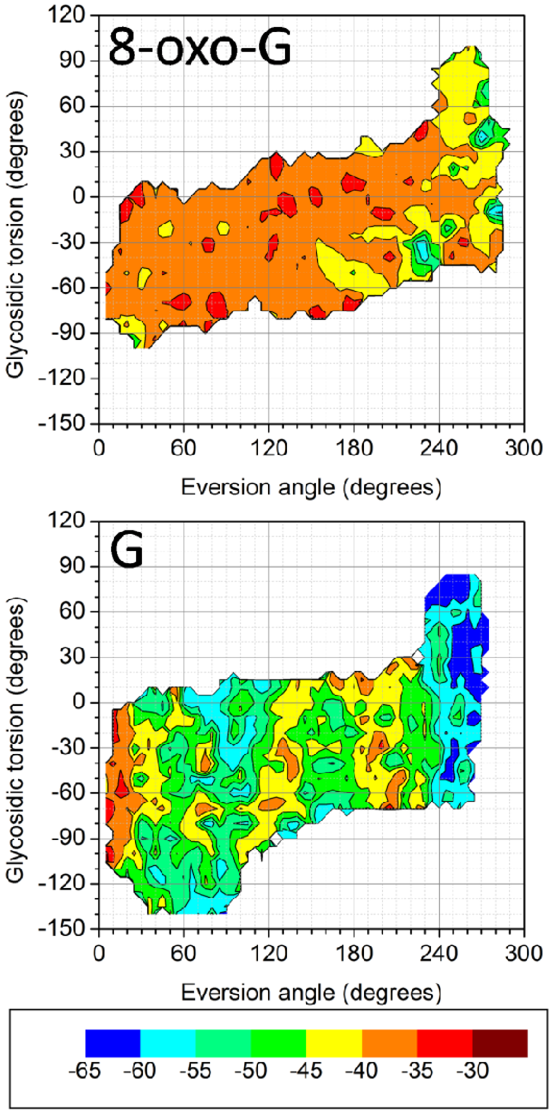


Figure 5-13. Interaction energy between the Lysine 257 and the second 5' phosphate from oxoG (top) and G (bottom). The legend is in kcal/mol, each contour represents 5 kcal/mol.

When oxoG contacts Arg 263 between an eversion angle of 20-100 degrees, an energetically favorable interaction occurs, Figure 5-11. The favorable interaction between Arg 263 and oxoG is caused by the amino group of Arg 263 interacting with the O8 of oxoG, Figure

5-14. A direct result of the energetically favorable interaction between Arg 263 to oxoG is that the Zinc finger distance to the DNA backbone is maintained similar to the intra-helical 2F5O structure (Figure 5-15) prohibiting Lys 257 from forming a hydrogen bond to the second 5' phosphate, discussed further below.

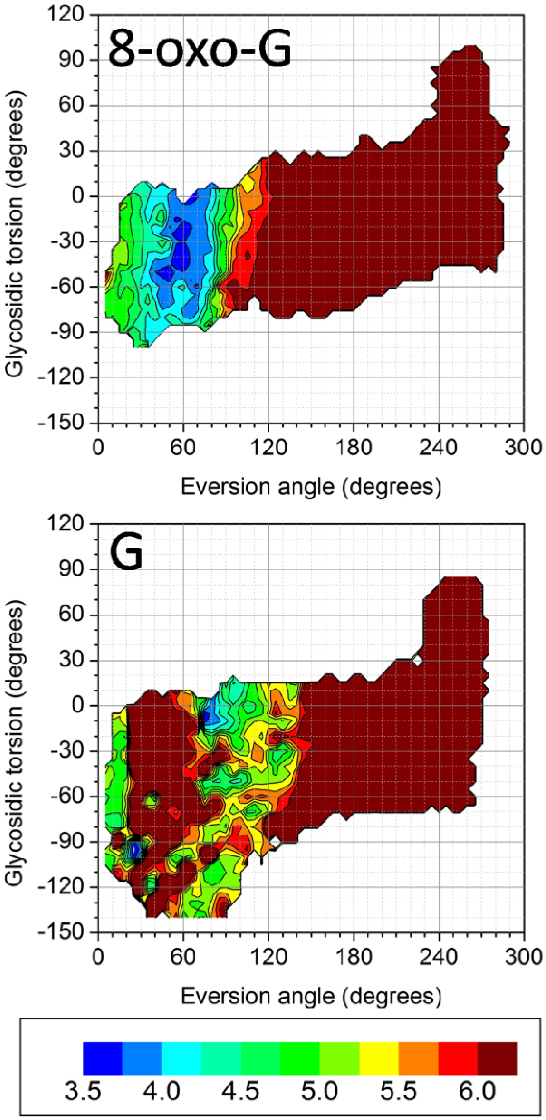


Figure 5-14. Distance measurement between the C8 of the everting nucleobase and the NH2 atom of Arginine 263 for oxoG (top) and G (bottom). The legend is in Angstroms and each contour represents .125 Å.

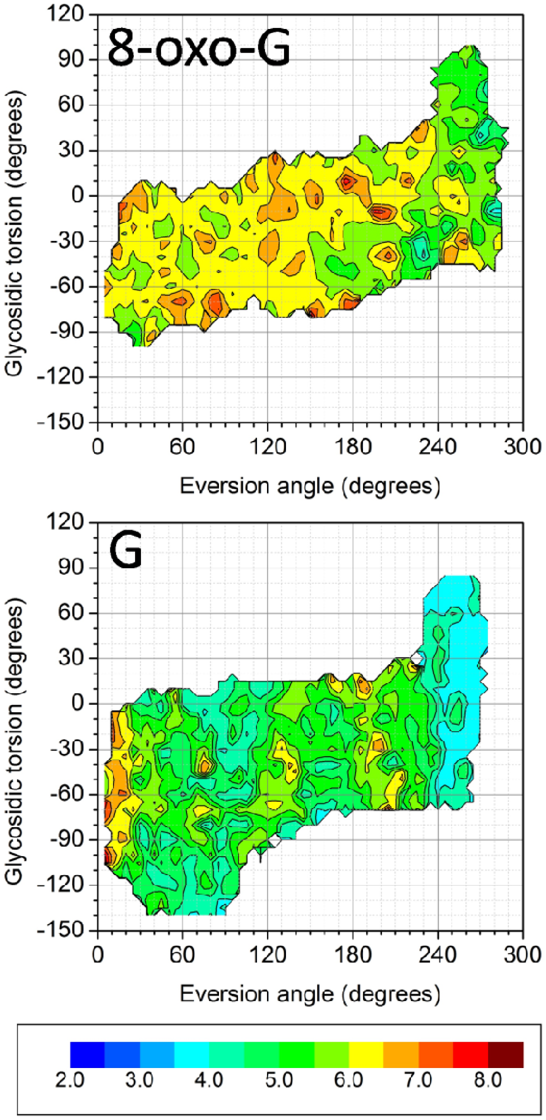


Figure 5-15. Distance measurement between the NZ of each Lysine 257 and the P atom of the second 5' phosphate for oxoG (top) and G (bottom). The legend is in Angstroms.

Through sequence conservation it was found that Arg 263 is largely conserved across all bacterial glycosylase phyla, giving further evidence of the important role of Arg 263 in Fpg's mechanism of base eversion (Figure 5-16). The only other residue that has been found to replace Arg 263 is a Lysine which is also positively charged and would be able to make the same interactions with the O8 of oxoG but not with G. The positively charged Lys 257 was also found to be conserved although in some phyla the functional residue is located 1 or 2 residues from the 257 position and/or replaced with another hydrogen bond donor.

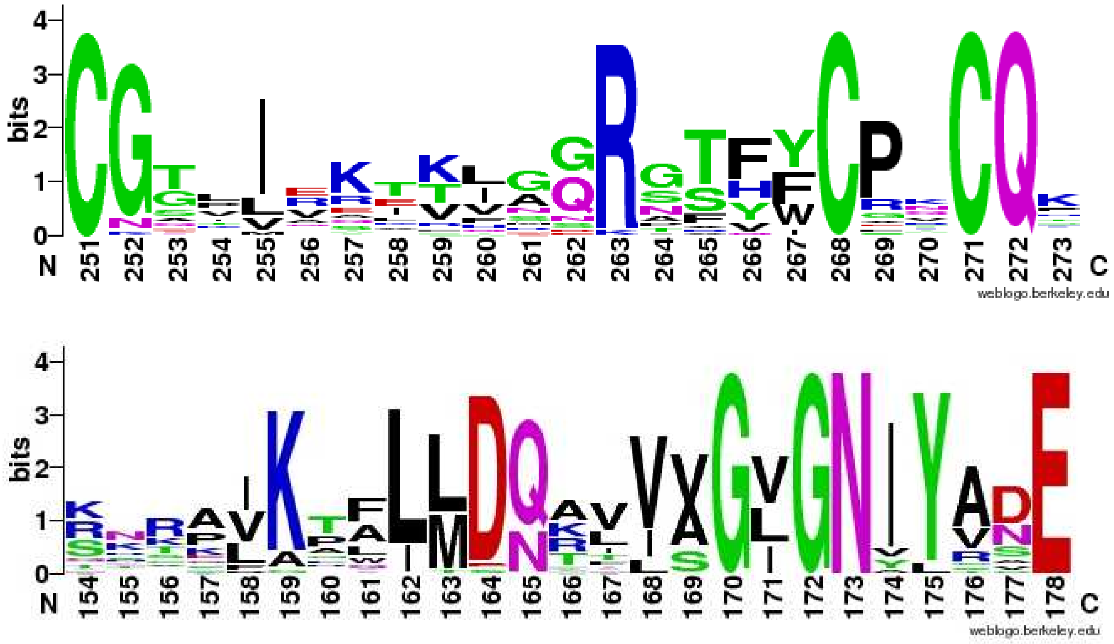


Figure 5-16. Sequence conservation of Arg263 and Gly264 region in Fpg, top. Sequence conservation in the region around Asn173, bottom.

5.3.2 Discrimination at the exo-site

Recognizing G in the exo-site

When inspecting the overall energetic pathways taken between the two systems, there is a stable region (exo-site) at an eversion angle of ~120 degrees for G and oxoG, Figure 5-6. For G to enter into this exo-site it would need to overcome an energetic barrier of ~7kcal/mol while oxoG would need to overcome a barrier of ~3kcal/mol. Fpg/MutM makes direct contacts with both G and oxoG in the exo-site, but these contacts are different, (Figure 5-17). As mentioned, Lys 257 forms a hydrogen bond with the second 5' phosphate group, therefore drawing the Zinc finger closer to the DNA backbone and the everting G, Figure 5-15. The early Lys 257 interaction with the phosphate prohibits G from interacting with the phosphate as a result of the close Zinc finger since Lys257 is already forming a strong interaction with it. When G enters the exo-site it forms two key interactions: one cation-pi stacking interaction with Arg 263, and also a discriminatory hydrogen bond between the backbone H of Gly 264 and the N7 atom on the imidazole ring of G (Figure 5-17, B). Both Arg 263 and Gly 264 are located on the tip of the Zinc finger of Fpg. The stacking interaction between Arg 263 and G are similar to the extrahelical stacking interaction between oxoG and Phe 319 in the active site of hOGG1 (PDB 1EBM).[184] In both stacking interactions the nucleobase is maintained by cation-pi or pi-pi stacking respectively. oxoG has a hydrogen on the analogous N7 atom of the imidazole ring, therefore precluding the possibility that a hydrogen bond would be able to form to Gly264 (Figure 5-2).

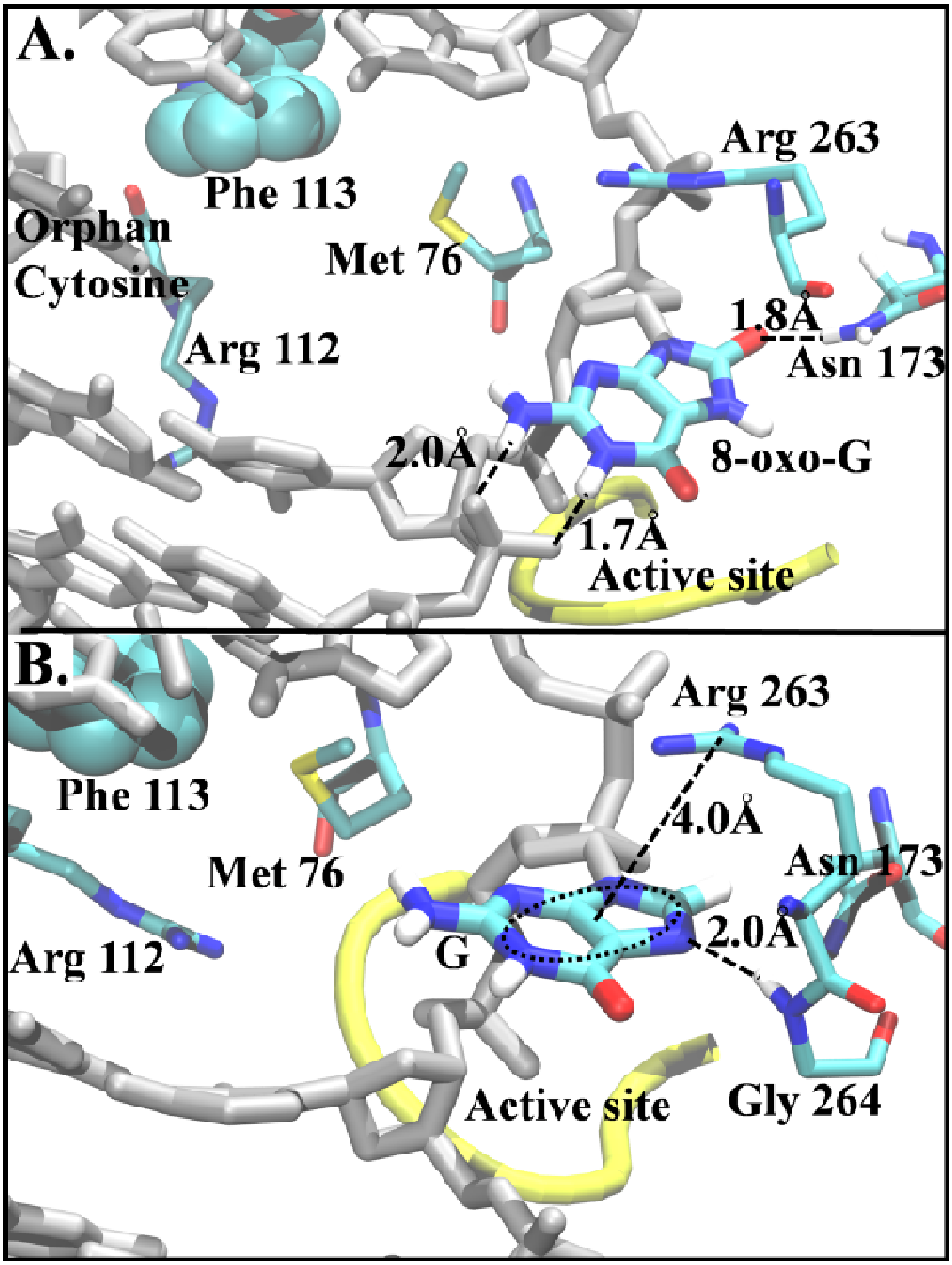


Figure 5-17. A.) Exo-site structure of 8-oxo-G at an eversion angle of 120 degrees and a glycosidic angle of -30 degrees. B.) Exo-site structure of G at an eversion angle of 100 degrees and a glycosidic angle of -30 degrees.

To ascertain how these specific interactions correspond to the free energies of eversion measured for G and oxoG in Figure 5-6, specific distances, along with their pair-wise interaction energies, were measured. The distances and interaction energies were then plotted with respect to the two reaction coordinates used in US, the eversion angle and the glycosidic angle (Figure 5-5). The cation-pi interaction was measured as the distance between the Arg 263 sidechain atoms (atoms CZ, NE, NH1 and NH2) and the heavy atoms of the purine ring of G and oxoG, (Figure 5-18. The cation-pi stacking that occurs between Arg 263 and G in the exo-site region (~120 degrees) does not occur for oxoG due to the Zinc finger being at a larger distance from the DNA backbone, Figure 5-15. To confirm that this interaction is an energetically favorable interaction for G in the exo-site, the pairwise interaction energy was separately measured between Arg 263 and each everting nucleobase (Figure 5-11). In the exo-site (~120° eversion angle) there is an energetically stabilizing interaction between G and Arg 263 which is not found to occur for oxoG confirming that this cation-pi interaction is stabilizing G in the exo-site (Figure 5-11). As mentioned earlier, a stabilizing interaction does occur between oxoG and Arg 263, but this happens before the exo-site is reached between an eversion angle of ~20 and ~80 degrees, due to Arg 263 interacting with the major groove face of oxoG (O8) as mentioned earlier and results in the blocking of Lys 257 from interacting with the second 5' phosphate group.

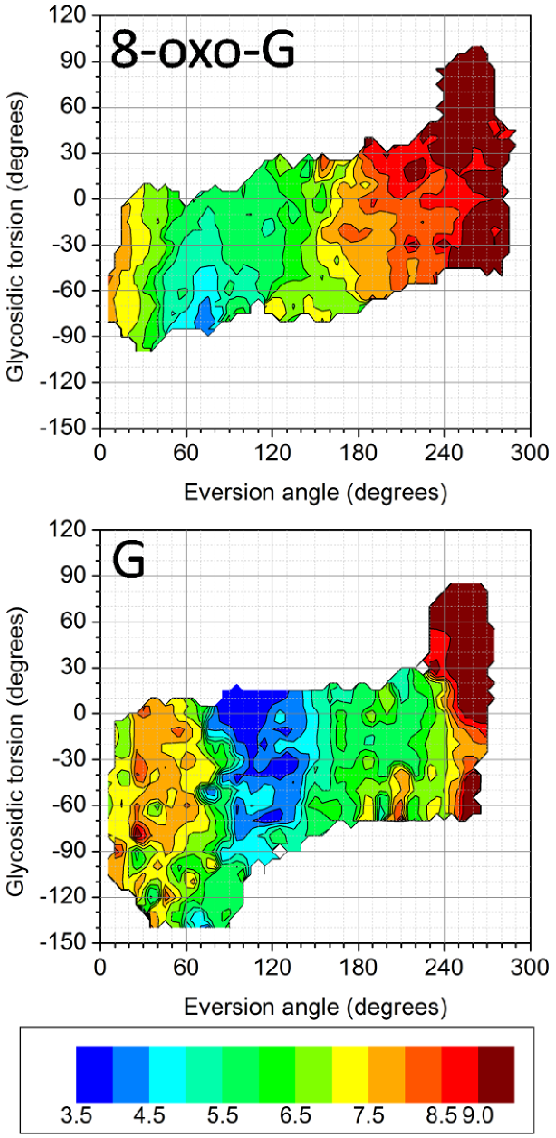


Figure 5-18. Cation-Pi stacking measurement. Distance measurement between the cyclic heavy atoms of the two rings on the purine ring and the heavy atoms of the side chain Arg263 (NH2, NH1, NE, CZ). The legend is in kcal/mol.

We hypothesized that the backbone of Gly 264 may be playing a role in discrimination between G and oxoG based on Figure 5-17. Interestingly it has been shown in hOGG1, Fpg's human functional analog, that the backbone of a glycine (Gly 42) is also used as a failsafe mechanism to discriminate between oxoG and G in the active site.[184] To address this hypothesis, the distances between the backbone H atom of Gly 264 and the N7 of G and oxoG were calculated, (Figure 5-2 and Figure 5-19). The distances confirm that at the exo-site a discriminatory hydrogen bond occurs between the H of Gly 264 and the N7 of G, but does not occur for oxoG. To measure whether this discriminatory hydrogen bond affects the energy of the system, the interaction energy between Gly 264 and each everting nucleobase was measured (Figure 5-20). The interaction energy shows that there is no substantial energetic difference between Gly 264 to G and oxoG at the exo-site suggesting that Gly264 is playing a role in recognition but not in the lowering of the free energy barrier to eversion for G. The sequence conservation analysis of Gly 264 across all bacterial glycosylase phyla did not show strict conservation (Figure 5-16). However, since the Gly 264 backbone H forms the hydrogen bond, any other residue, with the exception of proline, may substitute the glycine and still successfully hydrogen bond. Proline has not been reported as a substitute at position 264, which supports our structural analysis.

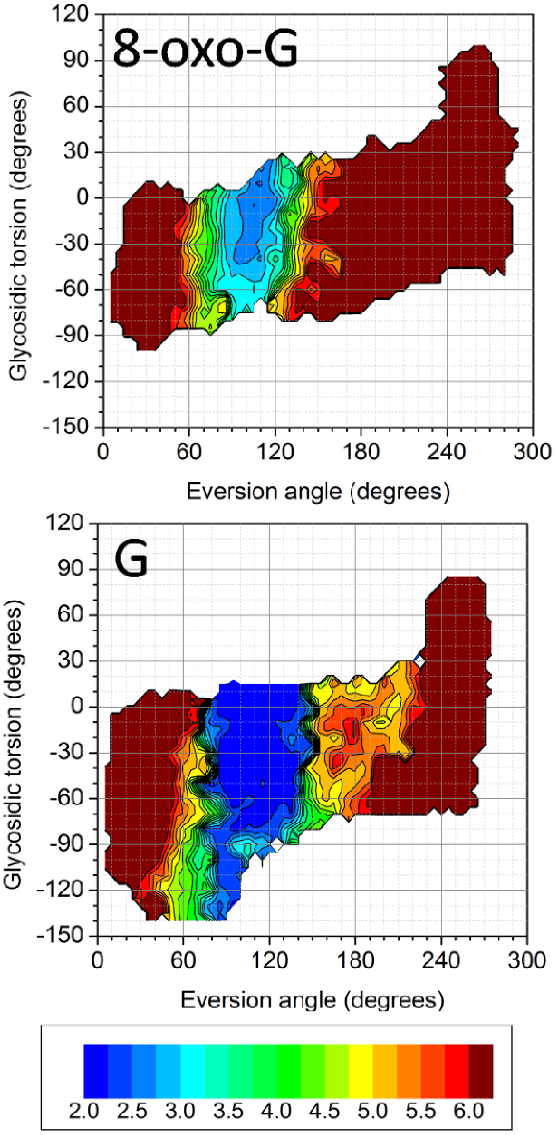


Figure 5-19. Distance measurement between the N7 of each flipping nucleobase and Gly264 backbone H for 8-oxo-G (top) and G (bottom). The legend is in kcal/mol.

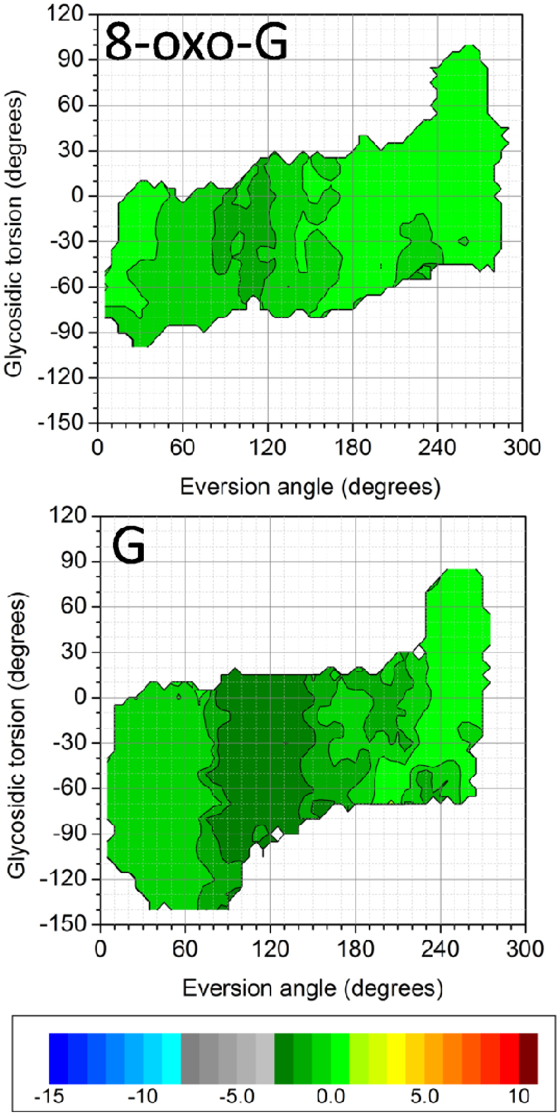


Figure 5-20. Energetic interaction between the flipping nucleobase and Gly264 for 8-oxo-G (top) and G (bottom). The legend is in kcal/mol.

Recognizing oxoG in the exo-site

Since oxoG does not form the hydrogen bond to Gly 264, oxoG is free to rotate away from Arg 263 and can contact the second 5' phosphate away from oxoG (Figure 5-17A). This interaction is able to occur since Lys 257 is not contacting the second 5' phosphate group as in G.wt. Two hydrogen bonds are made between the Watson-Crick face of oxoG and the second 5' phosphate group in the Fpg/MutM exo-site: one between the H21 of oxoG and the O2P of the phosphate group, and the other between H1 of oxoG and the O1P of the same phosphate group (Figure 5-17, A). OxoG is also able to form another lesion-specific hydrogen bond between its O8 and the amino group of Asn 173 (Figure 5-17, A).

To understand if the phosphate interactions only occur for oxoG in the exo-site, two distance measurements were made for the second 5' phosphate contact: the O1P atom of the phosphate to the N1 atom on the everting nucleobase, and the O2P atom of the phosphate to the N2 atom on the everting nucleobase (Figure 5-21 and Figure 5-22 respectively). The two hydrogen bonds that form for oxoG in the exo-site do not form for G in the exo-site. To confirm that these interactions contributed favorably for oxoG as opposed to G in the exo-site, pairwise energy decomposition was measured (Figure 5-23). The interaction energies for oxoG and G confirmed that oxoG is stabilized in the exo-site by interacting with the second 5' phosphate group, while G does not thus lowering the energetic barrier to eversion for oxoG (Figure 5-6). G cannot benefit from interacting with the second 5' phosphate group during eversion since Lys 257 is making a favorable contact with it, Figure 5-12 and Figure 5-13. Since oxoG can form hydrogen bonds to the second 5’ phosphate, the other side of the nucleobase can interact with Asn 173 in a favorable lesion-specific interaction.

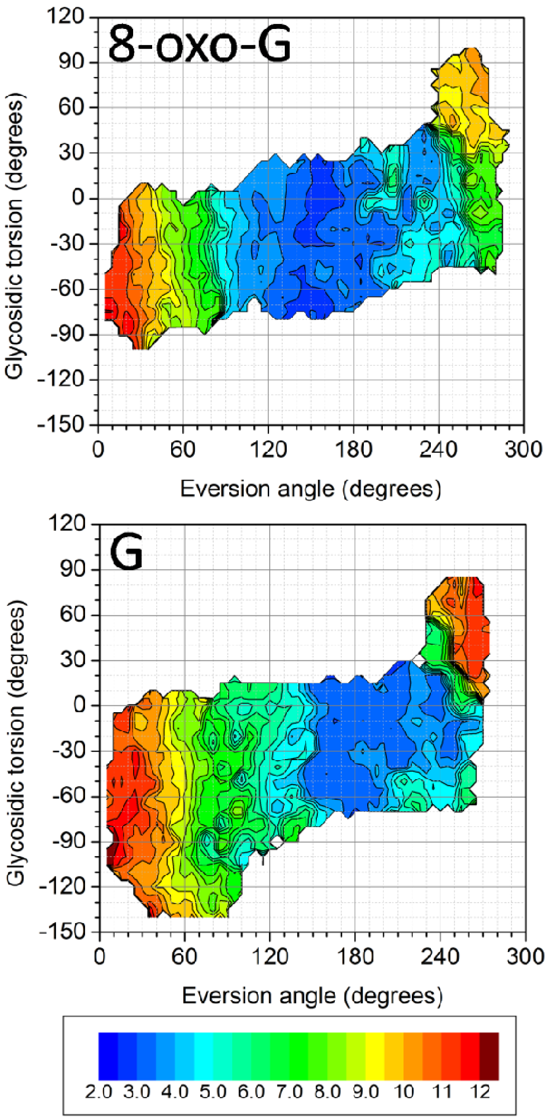


Figure 5-21. Distance measurements for O1P atom of the second 5' phosphate group away from the everting nucleobase to N1 of the everting nucleobase for 8-oxo-G and G (top and bottom respectively). The legend is in kcal/mol.

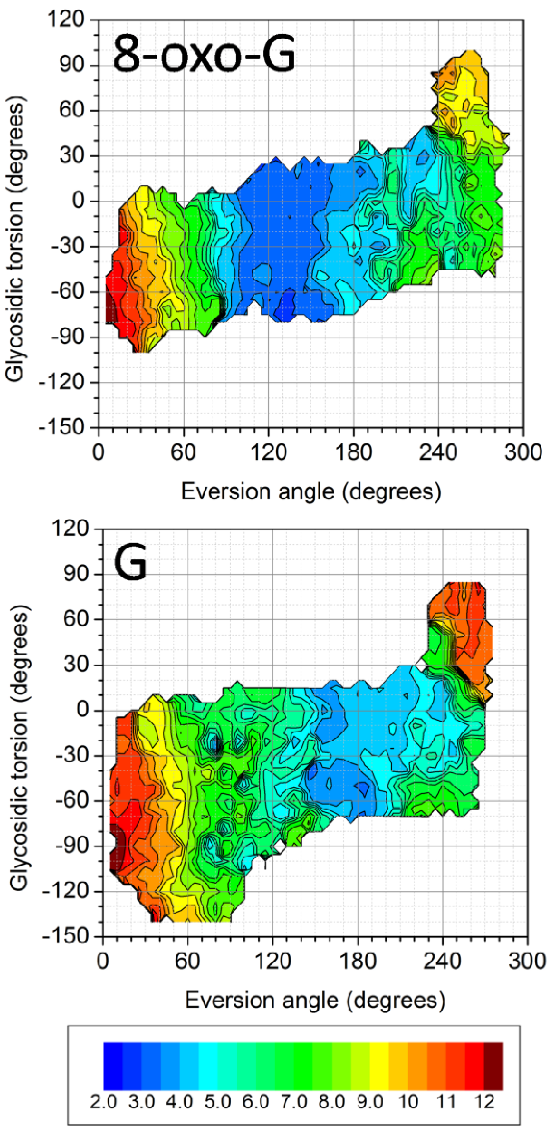


Figure 5-22. Distance measurements between O2P atom of the second 5' phosphate group away from the everting nucleobase to N2 atom of the everting nucleobase for 8-oxo-G and G (top and bottom respectively). The legend is in kcal/mol.

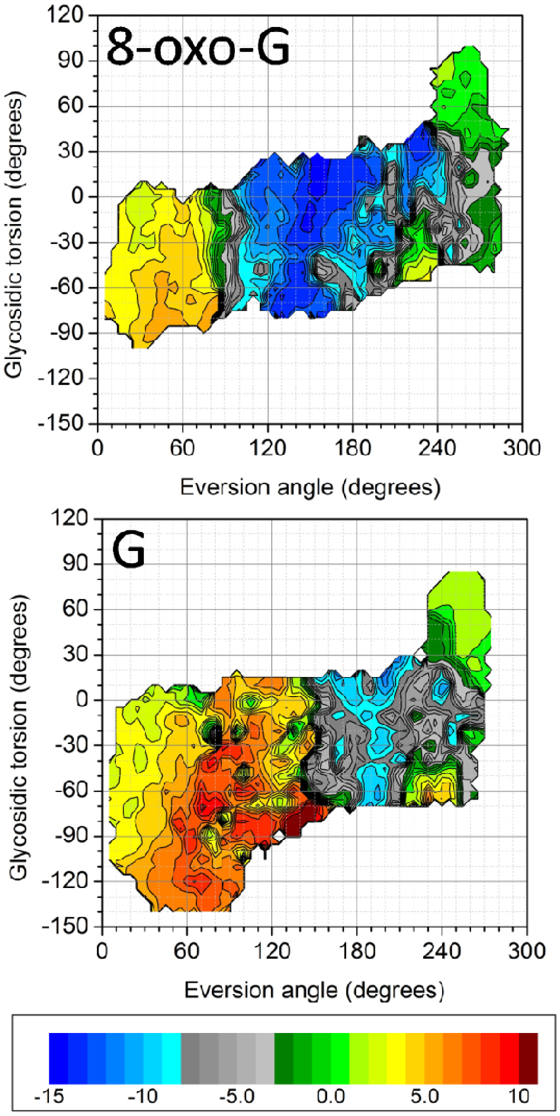


Figure 5-23. Interaction energy between the everting nucleobase and the second 5' phosphate group for 8-oxo-G and G (top and bottom respectively). The interaction energy was calculated between all atoms on the nucleobase to all atoms on the second 5' phosphate group (O1P, O2P, P, O3' and O5'). The legend is in kcal/mol.

To better understand if oxoG makes a contact in the exo-site with Asn 173, the distance of C8 of the everting nucleobase was measured to the ND2 of Asn 173 (Figure 5-24). Based on this distance measurement, it appeared that both oxoG and G come into close contact with the ND2 of Asn 173. Upon further structural inspection it was revealed that Asn 173 rotates below G in the exo-site to avoid contacting the electropositive H8 and does not appear to make a hydrogen bond, Figure 5-17. Therefore, the specific Asn 173/oxoG hydrogen bond may help to stabilize oxoG in the exo-site based on the interaction energy. To confirm this hypothesis, the interaction energy between the everting nucleobase and Asn 173 was measured (Figure 5-25). The interaction energies of G and oxoG confirmed that Asn 173 is stabilizing oxoG at the exo-site while giving no significant contribution to G at the exo-site. Sequence conservation was examined to confirm that Asn 173, a residue that we are hypothesizing plays a pivotal role in discrimination, is conserved (Figure 5-16). Asn173 does contact the DNA backbone in both the 1R2Y and 2F5O crystal structures.

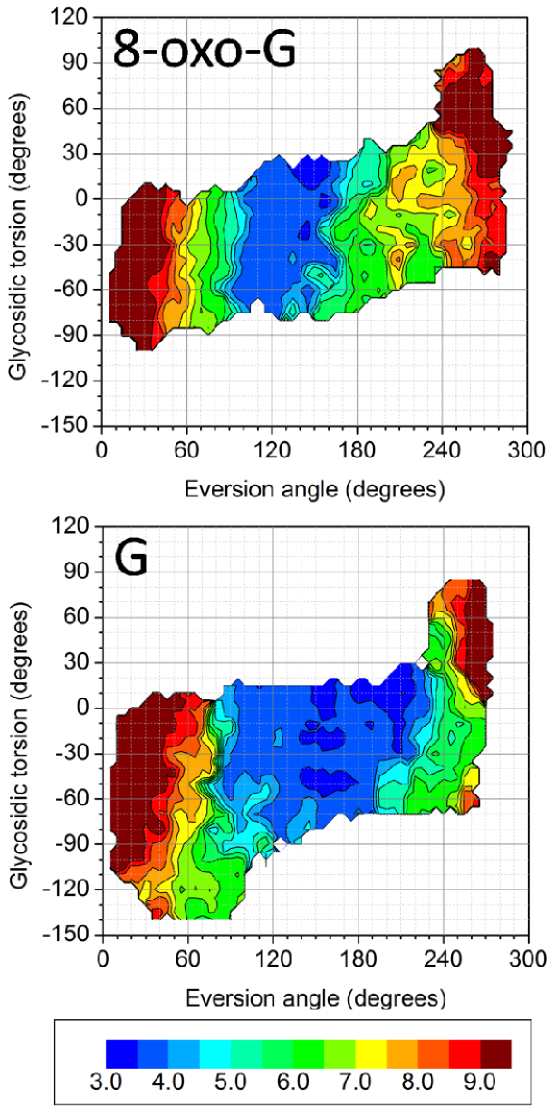


Figure 5-24. Distance measurement between the C8 atom of the everting nucleobase and the ND2 atom of Asn173 for 8-oxo-G and G (top and bottom respectively). The legend is in Angstroms.

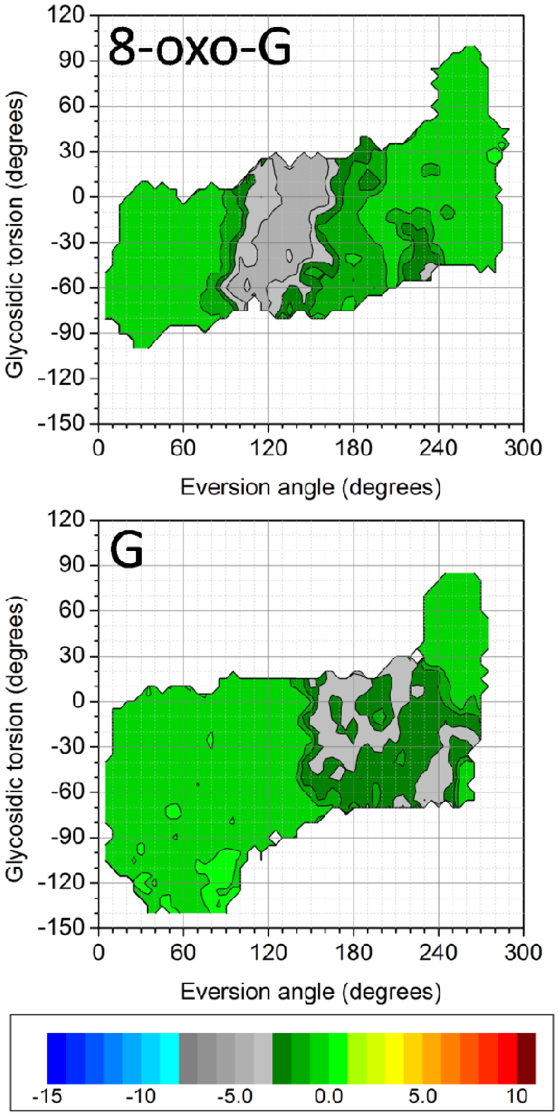


Figure 5-25. Interaction energy between the everting nucleobase and Asn173 for 8-oxo-G and G (top and bottom respectively). The legend is in kcal/mol.

Since Asn 173 appears to play a significant role in distinguishing between G and oxoG in the exo-site, we hypothesized that removing the specific interactions between the amino group of Asn 173 and the everting nucleobase would change the overall free energy barrier to eversion in the exo-site region of oxoG. To accomplish this, the amino group charges on Asn 173 were removed and umbrella sampling was rerun to obtain new PMF profiles (Figure 5-26). It was found that the exo-site for oxoG had significant increases as compared to the original PMF profiles (Figure 5-6 and Figure 5-26). This increase in energy was attributed to the loss of the key interaction between Asn 173 and oxoG. The energy of the extra-helical position also increased due to the loss of contact of Asn 173 with the 3' phosphate group, which appears to help stabilize oxoG when in the active site.

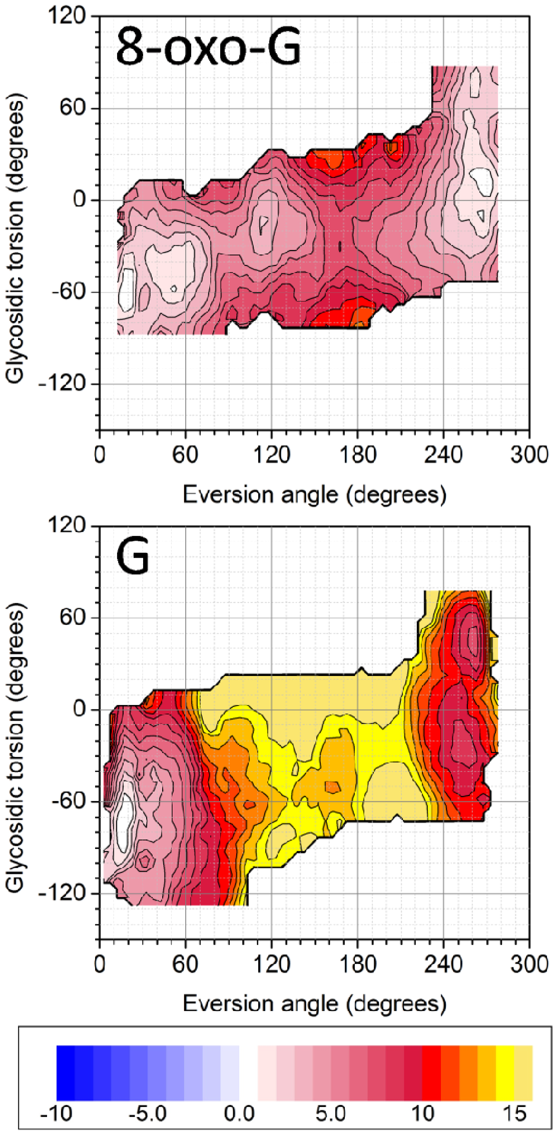


Figure 5-26. PMF profiles of 8-oxo-G (top) and G (bottom) where the charges on NH2 of Asn173 are turned off. The legend is in kcal/mol.

5.3.3 Validating that eversion occurs for nucleobases through fluorescence

To better understand the conformational transitions occurring for both G and oxoG when complexed to Fpg as suggested by our energy profiles (Figure 5-6), we designed a series of pre-steady-state kinetic studies using time course Trp fluorescence experiments on the *Eco*-Fpg-F113W system. We have shown through unrestrained molecular dynamics simulations (20ns) that the Fpg F113W mutation does not significantly affect wedge insertion (Figure 5-27, top**)**. This fluorescence sensitive mutation of the intercalating wedge F113W at the interrogation site will be sensitive to conformational changes, such as different extents of eversion, entrance into a transient stable exo-site, and entrance into the active site. A series of conformational changes for both systems was found to occur (Figure 5-27, middle and bottom). After *Eco*-Fpg-F113W binds to G, several conformational changes occur after initial wedge insertion, everting the undamaged nucleobase considerably but not completely into the active site for excision (Figure 5-27, middle). oxoG, undergoes similar conformational changes, as seen in the region from 0 to 0.1 sec of Figure 5-27, but these changes continue well past the 0.1 sec mark until excision is complete. Our experimental data is consistent with our energetic model, suggesting that both systems may enter our proposed exo-site but only oxoG continues everting. This data is also consistent with previous pre-steady-state kinetic studies.[74]

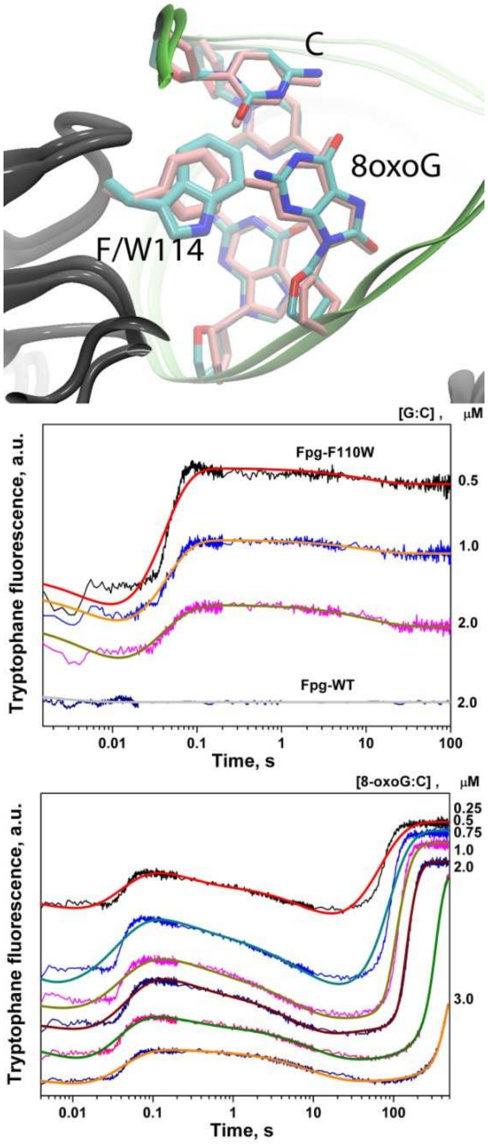


Figure 5-27. Top.) Overlap of structures from simulation of Bst-Fpg-DNA complex with a wild-type Phe113 wedge (pink) and a mutant Trp 113 wedge (colored by atom). Structure is shown only for the wedge and base pairs at the insertion site. Middle.) Experimental (jagged traces) and fit (smooth curves) time courses of Trp fluorescence changes during cleavage of the 8-oxo-G:C substrate by *Eco*-Fpg-F113W. Bottom.) Experimental (jagged traces) and fit (smooth curves) time courses of Trp fluorescence changes during binding of *Eco*-Fpg-F113W and wild-type *Eco*-Fpg to undamaged DNA. A.u., arbitrary units.

5.3.4 Leaving the exo-site and entering the active site

When leaving the exo-site both nucleobases interact with Fpg/MutM differently. For G the Lys257 interaction with the second 5’ phosphate group breaks and the nucleobase can interact with the second 5’ phosphate group, Figure 5-12 and Figure 5-23. This late interaction between G and the phosphate group occurs too late to benefit from any significant lowering of the energy barrier as seen in oxoG as in the exo-site. When oxoG leaves the exo-site (eversion angle of ~180 degrees) it loses the two hydrogen bonds between the second 5’ phosphate group which contributes to a slight energetic barrier on the energy landscape but makes an additional hydrogen bond to the phosphate group maintaining the low barrier, Figure 5-6.

5.4 *Conclusions*

In conclusion, Fpg/MutM facilitates base pair opening of both G and oxoG at no significant energetic cost through the major groove pathway. Early interactions with Arg263 specifically recognize the damaged side of oxoG and would accommodate other Fpg/MutM substrates such as FaPy-A and FaPy-G since they both contain the O8 atom in the identical position. The proposed exo-site mechanism where oxoG interacts with the second 5’ phosphate group as well as Asn173 contributes to the lowering of the overall energetic barrier to eversion for oxoG and would also accommodate the other two FaPy substrates. Our proposed Fpg/MutM mechanism for discrimination through the major groove draws several parallels to hOGG1 since base eversion it thought to occur through the major groove pathway and discrimination has been hypothesized to occur at the exo-site. Our minor groove energy profiles are consistent with previous work done on the Fpg/MutM system but the major groove pathway appears to be the dominant one where base eversion occurs. It remains to be seen if other glycosylases have similar mechanisms to eversion and discrimination.

Chapter 6 Future plans and concluding remarks

In Chapter 2, a combination of experimental and computational techniques were used to show that the nitrogen mustard structural analog proposed to study interstrand crosslink repair was in fact an accurate analog of nitrogen mustard. From our computations, we were able to show that the naturally occurring 1,3 interstrand crosslink formed from nitrogen mustard gave similar amounts of distortion as compared to that of the analog. In the near future, two major steps can be taken to improve this type of analysis. First, running simulations longer than 50 ns may help give better statistics on the damage. This result, however, is highly dependent on the forcefield parameters of DNA that other research groups are currently working to improve.[81] Another step towards improved modeling of 1,3 interstrand crosslinks can be taken by parameterizing and simulating longer and shorter crosslinks to see if varying the length and charge has a significant effect on distortion. If there is a relationship between distortion and linker length, then it would be beneficial to have a way to predict distortion computationally and be able to measure the effects of larger or smaller amounts of distortion through experiment.[113,203]

A metric to quantify base eversion was discussed at length in Chapter 3. Two similar ways to measure base eversion were proposed, where the only variation was in points 2 and 3 on the dihedral angle (either the flanking sugars or phosphate groups). The proposed eversion angle was then used in Chapters 4 and 5 to help quantify the free energy required for eversion in various environments when bound to Fpg. What was not discussed was in what circumstances it is beneficial to choose one measure over the other (i.e., flanking sugars or the phosphate groups) as points 2 and 3 of the PHI dihedral. When trying to evert a nucleobase in a complex system where steric clashes will occur during the eversion process, we found it best to select the phosphate groups for the eversion measurement, since this will give a more drawn out eversion angle, thus giving more of an opportunity to separate different events. Because the eversion process takes longer when the phosphates are selected over the sugars due to the greater sensitivity of the angle of rotation, this allows a greater resolution of the process. If one were to measure eversion in duplex DNA, we have found it best to select the sugar groups for the measurement, as everting a nucleobase in naked DNA causes some local distortion and selecting the sugars over the phosphate groups is stabilized by the wider points 2 and 3. Another important point to consider is whether the sugars or phosphates are playing some role in the process of interest. If so, then careful attention should be made not to create an unintentional biasing.

In several chapters I have discussed coupling partial nudged elastic band method with umbrella sampling to obtain free energy pathways to eversion. One caveat that was not discussed in these chapters was the simulation length for the windows in umbrella sampling. Allowing the windows to equilibrate is important, but it is also important not to simulate the windows for too long, thus creating artificial distortion. When restraints are applied to all windows in umbrella sampling, these restraints will invariably be applied to some regions that are of higher energy; thus, in those regions, the system will begin to drift away from the high energy regions even if the reaction coordinate(s) do not show this effect. This drifting is caused by the system adjusting to the reaction coordinate being restrained to a region of high energy and it not being able to escape it. In many cases in our lab, we use structures that were extracted from partial nudged elastic band method for umbrella sampling. These umbrella sampling simulations need to be representative simulations of the PNEB pathway and any drifting that occurs due to extremely long window length can create artificial results. In one example, we found that the RMSD of the first half and second half of the umbrella sampling data were substantially different. In this example, there were many high-energy regions in the pathway. When we ran separate WHAM analysis we found the PMFs to be different as well.

Chapter Four discusses the consequences of removing the aromatic Phe 113 wedge from Fpg. It is discussed in this Chapter in great detail that removing the wedge stabilizes the intrahelical position for the F113A mutant while keeping the wedge destabilizes the intrahelical region thus facilitating the eversion process. This Chapter also discusses how eversion naturally occurs for abasic sites through the major groove in duplex DNA. It would be interesting to simulate Fpg bound to DNA containing an abasic site to see whether eversion occurs through the major or minor groove and if so, what interactions are formed when this happens.

In Chapter 5, we focus on comparing the structural differences of G and 8-oxo-G as eversion occurs when bound by Fpg. Fpg is also known to correct other less common oxidative lesions such as Fapy-G, Fapy-A and 5-OH-C.[187,188] It would be interesting to computationally study and compare 8-oxo-G/abasic sites/Fapy-G/Fapy-A and 5-OH-C in a similar manner as done in Chapter 5.[187,188] The Fapy lesions are open purine rings and one would therfore expect that these flexible lesions (Fapy-G and Fapy-A) would be more promiscuous with their interactions along the eversion path.

Several chapters have covered the function of the Fpg glycosylase. These chapters did not address the directions in which our modeling efforts of Fpg are headed. In general, the biological community has great interest in damage recognition and how discrimination occurs in the BER glycosylases. In this work, we have explored the steps that occur after damage recognition, but did not address the issue of actual recognition. Several group members are currently exploring this issue by measuring translocation of Fpg along damaged and undamaged DNA in a similar approach as discussed in Chapter 5, where PNEB is used to uncover the pathway while umbrella sampling is used to obtain the free energy along the process. This work will be very valuable to the computational and experimental community. It would also be interesting to couple translocation and the eversion process of Fpg to obtain atomic resolution of these processes together.

References

1. Friedberg, E.C. et al. DNA Repair and Mutagensis. *ASM Press* **Washington, DC**(2006).

2. Dong, H. et al. Mutagenic potential of benzo[a] pyrene-derived DNA adducts positioned in codon 273 of the human P53 gene. *Biochemistry* **43**, 15922-15928 (2004).

3. Grollman, A.P. et al. Aristolochic acid and the etiology of endemic (Balkan) nephropathy. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 12129-12134 (2007).

4. Bjelland, S. & Seeberg, E. Mutagenicity, toxicity and repair of DNA base damage induced by oxidation. *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis* **531**, 37-80 (2003).

5. Gutteridge, J.M.C. & Halliwell, B. Free radicals and antioxidants in the year 2000 - A historical look to the future. *Reactive Oxygen Species: From Radiation to Molecular Biology* **899**, 136-147 (2000).

6. Shangary, S. & Wang, S.M. Targeting the MDM2-p53 interaction for cancer therapy. *Clinical Cancer Research* **14**, 5318-5324 (2008).

7. Barakat, K., Mane, J., Friesen, D. & Tuszynski, J. Ensemble-based virtual screening reveals dual-inhibitors for the p53-MDM2/MDMX interactions. *Journal of Molecular Graphics & Modelling* **28**, 555-568 (2010).

8. Scharer, O.D. DNA interstrand crosslinks: natural and drug-induced DNA adducts that induce unique cellular responses. *Chembiochem* **6**, 27-32 (2005).

9. Noll, D.M., Mason, T.M. & Miller, P.S. Formation and repair of interstrand cross-links in DNA. *Chemical Reviews* **106**, 277-301 (2006).

10. O'Connor, P.M. & Kohn, K.W. Comparative pharmacokinetics of DNA lesion formation and removal following treatment of L1210 cells with nitrogen mustards. *Cancer Commun* **2**, 387-94 (1990).

11. Akkari, Y.M.N., Bateman, R.L., Reifsteck, C.A., Olson, S.B. & Grompe, M. DNA replication is required to elicit cellular responses to psoralen-induced DNA interstrand cross-links. *Molecular and Cellular Biology* **20**, 8283-8289 (2000).

12. Kasai, H., Hayami, H. & Nishimura, S. Hydroxylation of Deoxyguanosine at the C-8 Position by Mutagens and Carcinogens. *Mutation Research* **130**, 366-366 (1984).

13. Dizdaroglu, M. Formation of an 8-Hydroxyguanine Moiety in Deoxyribonucleic-Acid on Gamma-Irradiation in Aqueous-Solution. *Biochemistry* **24**, 4476-4481 (1985).

14. McCullough, A.K., Dodson, M.L. & Lloyd, R.S. Initiation of base excision repair: Glycosylase mechanisms and structures. *Annual Review of Biochemistry* **68**, 255-285 (1999).

15. Watson, J.D. & Crick, F.H.C. Molecular Structure of Nucleic Acids - a Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737-738 (1953).

16. Fukui, K. DNA mismatch repair in eukaryotes and bacteria. *J Nucleic Acids* **2010**(2010).

17. Dixit, S.B. et al. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: Sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophysical Journal* **89**, 3721-3740 (2005).

18. Dickerson, R.E. DNA bending: The prevalence of kinkiness and the virtues of normality. *Nucleic Acids Research* **26**, 1906-1926 (1998).

19. Yonetani, Y. & Kono, H. Sequence Dependencies of DNA Deformability and Hydration in the Minor Groove. *Biophysical Journal* **97**, 1138-1147 (2009).

20. Bonvin, A.M.J.J., Sunnerhagen, M., Otting, G. & van Gunsteren, W.F. Water molecules in DNA recognition II: A molecular dynamics view of the structure and hydration of the trp operator. *Journal of Molecular Biology* **282**, 859-873 (1998).

21. Wu, B., Mohideen-Abdul, K., Vasudevan, D. & Davey, C.A. Structural Insight into the Sequence Dependence of Nucleosome Positioning. *Structure* **18**, 528-536 (2010).

22. Sarai, A., Mazur, J., Nussinov, R. & Jernigan, R.L. Origin of DNA Helical Structure and Its Sequence Dependence. *Biophysical Journal* **53**, A306-A306 (1988).

23. Wing, R. et al. Crystal-Structure Analysis of a Complete Turn of B-DNA. *Nature* **287**, 755-758 (1980).

24. Freisinger, E., Fernandes, A., Grollman, A.P. & Kisker, C. Crystallographic characterization of an exocyclic DNA adduct: 3,N-4-etheno-2 '-deoxycytidine in the dodecamer 5 '-CGCGAATT epsilon CGCG-3 '. *Journal of Molecular Biology* **329**, 685-697 (2003).

25. Cullinan, D., Johnson, F., Grollman, A.P., Eisenberg, M. & delosSantos, C. Solution structure of a DNA duplex containing the exocyclic lesion 3,N-4-etheno-2'-deoxycytidine opposite 2'-deoxyguanosine. *Biochemistry* **36**, 11933-11943 (1997).

26. Korobka, A. et al. Solution structure of an oligodeoxynucleotide duplex containing the exocyclic lesion 3,N-4-etheno-2'-deoxycytidine opposite 2'-deoxyadenosine, determined by NMR spectroscopy and restrained molecular dynamics. *Biochemistry* **35**, 13310-13318 (1996).

27. Case, D.A. et al. The Amber biomolecular simulation programs. *Journal of Computational Chemistry* **26**, 1668-1688 (2005).

28. Coin, F., Oksenych, V. & Egly, J.M. Distinct roles for the XPB/p52 and XPD/p44 subcomplexes of TFIIH in damaged DNA opening during nucleotide excision repair. *Molecular Cell* **26**, 245-256 (2007).

29. Fan, L. et al. Conserved XPB core structure and motifs for DNA unwinding: Implications for pathway selection of transcription or excision repair. *Molecular Cell* **22**, 27-37 (2006).

30. Gillet, L.C.J. & Scharer, O.D. Molecular mechanisms of mammalian global genome nucleotide excision repair. *Chemical Reviews* **106**, 253-276 (2006).

31. Porecha, R.H. & Stivers, J.T. Uracil DNA glycosylase uses DNA hopping and short-range sliding to trap extrahelical uracils. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 10791-10796 (2008).

32. Blainey, P.C., van Oijent, 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**, 5752-5757 (2006).

33. Blainey, P.C., van Oijen, A.M., King, B.H., Verdine, G.L. & Xie, X.S. Tracking single DNA repair proteins on stretched DNA by fluorescence imaging. *Abstracts of Papers of the American Chemical Society* **228**, U299-U299 (2004).

34. Blainey, P.C. et al. Nonspecifically bound proteins spin while diffusing along DNA. *Nature Structural & Molecular Biology* **16**, 1224-U34 (2009).

35. Doherty, A.J., Serpell, L.C. & Ponting, C.P. The helix-hairpin-helix DNA-binding motif: A structural basis for non-sequence-specific recognition of DNA. *Nucleic Acids Research* **24**, 2488-2497 (1996).

36. Eichman, B.F., O'Rourke, E.J., Radicella, J.P. & Ellenberger, T. Crystal structures of 3-methyladenine DNA glycosylase MagIII and the recognition of alkylated bases. *Embo Journal* **22**, 4898-4909 (2003).

37. Guan, Y. et al. MutY catalytic core, mutant and bound adenine structures define specificity for DNA repair enzyme superfamily. *Nature Structural Biology* **5**, 1058-1064 (1998).

38. Kuo, C.F. et al. Atomic-Structure of the DNA-Repair [4fe-4s] Enzyme Endonuclease-Iii. *Science* **258**, 434-440 (1992).

39. Labahn, J. et al. Structural basis for the excision repair of alkylation-damaged DNA. *Cell* **86**, 321-329 (1996).

40. Mol, C.D., Arvai, A.S., Begley, T.J., Cunningham, R.P. & Tainer, J.A. Structure and activity of a thermostable thymine-DNA glycosylase: evidence for base twisting to remove mismatched normal DNA bases. *Journal of Molecular Biology* **315**, 373-384 (2002).

41. Yamagata, Y. et al. Three-dimensional structure of a DNA repair enzyme, 3-methyladenine DNA glycosylase II, from Escherichia coli. *Cell* **86**, 311-319 (1996).

42. Tsodikov, O.V. et al. Structural basis for the recruitment of ERCC1-XPF to nucleotide excision repair complexes by XPA. *EMBO J* **26**, 4768-76 (2007).

43. 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**, 141-160 (2002).

44. Giudice, E., Varnai, P. & Lavery, R. Base pair opening within B-DNA: free energy pathways for GC and AT pairs from umbrella sampling simulations (vol 31, pg 1434, 2003). *Nucleic Acids Research* **31**, 2703-2703 (2003).

45. Giudice, E., Varnai, P. & Lavery, R. Energetic and conformational aspects of A : T base-pair opening within the DNA double helix. *Chemphyschem* **2**, 673-+ (2001).

46. Giudice, E. & Lavery, R. Nucleic acid base pair dynamics: The impact of sequence and structure using free-energy calculations. *Journal of the American Chemical Society* **125**, 4998-4999 (2003).

47. 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**, 68-73 (2003).

48. Huang, N. & MacKerell, A.D. Specificity in protein-DNA interactions: Energetic recognition by the (cytosine-C5)-methyltransferase from HhaI. *Journal of Molecular Biology* **345**, 265-274 (2005).

49. Priyakumar, U.D. & MacKerell, A.D. Computational approaches for investigating base flipping in oligonucleotides. *Chemical Reviews* **106**, 489-505 (2006).

50. Priyakumar, U.D. & MacKerell, A.D. Base flipping in a GCGC containing DNA dodecamer: A comparative study of the performance of the nucleic acid force fields, CHARMM, AMBER, and BMS. *Journal of Chemical Theory and Computation* **2**, 187-200 (2006).

51. Varnai, P. & Lavery, R. Base flipping in DNA: Pathways and energetics studied with molecular dynamic simulations. *Journal of the American Chemical Society* **124**, 7272-7273 (2002).

52. Zheng, H. et al. Base Flipping Free Energy Profiles for Damaged and Undamaged DNA. *Chemical Research in Toxicology* **23**, 1868-1870 (2010).

53. Chen, Y.Z., Mohan, V. & Griffey, R.H. The opening of a single base without perturbations of neighboring nucleotides: A study on crystal B-DNA duplex d(CGCGAATTCGCG)(2). *Journal of Biomolecular Structure & Dynamics* **15**, 765-+ (1998).

54. Keepers, J., Kollman, P.A. & James, T.L. Molecular Mechanical Studies of Base-Pair Opening in D(Cgcgc)-D(Gcgcg), Dg5.Dc5, D(Tatat)-D(Atata), and Da5.Dt5 in the B-Form and Z-Form of DNA. *Biopolymers* **23**, 2499-2511 (1984).

55. Keepers, J.W., Kollman, P.A., Weiner, P.K. & James, T.L. Molecular Mechanical Studies of DNA Flexibility - Coupled Backbone Torsion Angles and Base-Pair Openings. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* **79**, 5537-5541 (1982).

56. Noll, D.M., Noronha, A.M., Wilds, C.J. & Miller, P.S. Preparation of interstrand cross-linked DNA oligonucleotide duplexes. *Frontiers in Bioscience* **9**, 421-437 (2004).

57. Angelov, T., Guainazzi, A. & Scharer, O.D. Generation of DNA Interstrand Cross-Links by Post-Synthetic Reductive Amination. *Organic Letters* **11**, 661-664 (2009).

58. Swenson, M.C., Paranawithana, S.R., Miller, P.S. & Kielkopf, C.L. Structure of a DNA repair substrate containing an alkyl interstrand cross-link at 1.65 angstrom resolution. *Biochemistry* **46**, 4545-4553 (2007).

59. Remias, M.G., Lee, C.S. & Haworth, I.S. Molecular-Dynamics Simulations of Chlorambucil DNA-Adducts - a Structural Basis for the 5'-Gnc Interstrand DNA Cross-Link Formed by Nitrogen Mustards. *Journal of Biomolecular Structure & Dynamics* **12**, 911-936 (1995).

60. Dong, Q. et al. A structural basis for a phosphoramide mustard-induced DNA interstrand cross-link at 5'-d(GAC). *Proceedings of the National Academy of Sciences of the United States of America* **92**, 12170-12174 (1995).

61. Rink, S.M. et al. Covalent Structure of a Nitrogen Mustard-Induced DNA Interstrand Cross-Link - an N7-to-N7 Linkage of Deoxyguanosine Residues at the Duplex Sequence 5'-D(Gnc). *Journal of the American Chemical Society* **115**, 2551-2557 (1993).

62. Song, K., Santos, C.D.L., Grollman, A.P. & Simmerling, C.L. Computational analysis of the lesion recognition mechanism of formamido-pyrimidine DNA glycosylase. *Chemical Research in Toxicology* **20**, 2000-2000 (2007).

63. Song, K., Hornak, V., Santos, C.D., Grollman, A.P. & Simmerling, C. Computational analysis of the mode of binding of 8-oxoguanine to formamidopyrimidine-DNA glycosylase. *Biochemistry* **45**, 10886-10894 (2006).

64. Cheng, X.L. et al. Dynamic behavior of DNA base pairs containing 8-oxoguanine. *Journal of the American Chemical Society* **127**, 13906-13918 (2005).

65. Zaika, E.I. et al. Substrate discrimination by formamidopyrimidine-DNA glycosylase - A mutational analysis. *Journal of Biological Chemistry* **279**, 4849-4861 (2004).

66. Gilboa, R. et al. Structure of formamidopyrimidine-DNA glycosylase covalently complexed to DNA. *Journal of Biological Chemistry* **277**, 19811-19816 (2002).

67. Zharkov, D.O., Rieger, R.A., Iden, C.R. & Grollman, A.P. NH2-terminal proline acts as a nucleophile in the glycosylase/AP-lyase reaction catalyzed by Escherichia coli formamidopyrimidine-DNA glycosylase (Fpg) protein. *Journal of Biological Chemistry* **272**, 5335-5341 (1997).

68. Tchou, J. & Grollman, A.P. The Catalytic Mechanism of Fpg Protein - Evidence for a Schiff-Base Intermediate and Amino-Terminus Localization of the Catalytic Site. *Journal of Biological Chemistry* **270**, 11671-11677 (1995).

69. Bhagwat, M. & Gerlt, J.A. 3'- and 5'-strand cleavage reactions catalyzed by the Fpg protein from Escherichia coli occur via successive beta- and delta-elimination mechanisms, respectively. *Biochemistry* **35**, 659-665 (1996).

70. Fromme, J.C. & Verdine, G.L. DNA lesion recognition by the bacterial repair enzyme MutM. *Journal of Biological Chemistry* **278**, 51543-8 (2003).

71. Banerjee, A., Santos, W.L. & Verdine, G.L. Structure of a DNA glycosylase searching for lesions. *Science* **311**, 1153-7 (2006).

72. Qi, Y. et al. Encounter and extrusion of an intrahelical lesion by a DNA repair enzyme. *Nature* **462**, 762-U79 (2009).

73. Qi, Y., Spong, M.C., Nam, K., Karplus, M. & Verdine, G.L. Entrapment and Structure of an Extrahelical Guanine Attempting to Enter the Active Site of a Bacterial DNA Glycosylase, MutM. *Journal of Biological Chemistry* **285**, 1468-1478 (2010).

74. Kuznetsov, N.A., Zharkov, D.O., Koval, V.V., Buckle, M. & Fedorova, O.S. Reversible chemical step and rate-limiting enzyme regeneration in the reaction catalyzed by formamidopyrimidine-DNA glycosylase. *Biochemistry* **48**, 11335-43 (2009).

75. Kuznetsov, N.A. et al. Pre-steady-state kinetic study of substrate specificity of Escherichia coli formamidopyrimidine-DNA glycosylase. *Biochemistry* **46**, 424-435 (2007).

76. Koval, V.V., Kuznetsov, N.A., Ishchenko, A.A., Saparbaev, M.K. & Fedorova, O.S. Real-time studies of conformational dynamics of the repair enzyme E. coli formamidopyrimidine-DNA glycosylase and its DNA complexes during catalytic cycle. *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis* **685**, 3-10 (2010).

77. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **14**, 33-& (1996).

78. Hornak, V. et al. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins-Structure Function and Bioinformatics* **65**, 712-725 (2006).

79. Cornell, W.D. et al. A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *Journal of the American Chemical Society* **117**, 5179-5197 (1995).

80. Weiner, S.J. et al. A New Force-Field for Molecular Mechanical Simulation of Nucleic-Acids and Proteins. *Journal of the American Chemical Society* **106**, 765-784 (1984).

81. Perez, A. et al. Refinenement of the AMBER force field for nucleic acids: Improving the description of alpha/gamma conformers. *Biophysical Journal* **92**, 3817-3829 (2007).

82. Wickstrom, L., Okur, A. & Simmerling, C. Evaluating the Performance of the ff99SB Force Field Based on NMR Scalar Coupling Data. *Biophysical Journal* **97**, 853-856 (2009).

83. Nucci, N.V., Pometun, M.S. & Wand, A.J. Site-resolved measurement of water-protein interactions by solution NMR. *Nature Structural & Molecular Biology* **18**, 245-9 (2011).

84. Makarov, V., Pettitt, B.M. & Feig, M. Solvation and hydration of proteins and mucleic acids: A theoretical view of simulation and experiment. *Accounts of Chemical Research* **35**, 376-384 (2002).

85. Bagchi, B. Water dynamics in the hydration layer around proteins and micelles. *Chemical Reviews* **105**, 3197-3219 (2005).

86. 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**, 926-935 (1983).

87. 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**, 10089-10092 (1993).

88. Cheatham, T.E., Miller, J.L., Fox, T., Darden, T.A. & Kollman, P.A. Molecular-Dynamics Simulations on Solvated Biomolecular Systems - the Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, Rna, and Proteins. *Journal of the American Chemical Society* **117**, 4193-4194 (1995).

89. 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**, 6127-6129 (1990).

90. Horton, J.R. et al. Caught in the act: visualization of an intermediate in the DNA base-flipping pathway induced by HhaI methyltransferase. *Nucleic Acids Research* **32**, 3877-3886 (2004).

91. Bergonzo, C., Campbell, A.J., Walker, R.C. & Simmerling, C. A Partial Nudged Elastic Band Implementation for Use With Large or Explicitly Solvated Systems. *International Journal of Quantum Chemistry* **109**, 3781-3790 (2009).

92. Schlitter, J., Engels, M., Kruger, P., Jacoby, E. & Wollmer, A. Targeted Molecular-Dynamics Simulation of Conformational Change - Application to the T[--]R Transition in Insulin. *Molecular Simulation* **10**, 291-& (1993).

93. Isralewitz, B., Gao, M. & Schulten, K. Steered molecular dynamics and mechanical functions of proteins. *Current Opinion in Structural Biology* **11**, 224-230 (2001).

94. Kumar, S., Bouzida, D., Swendsen, R.H., Kollman, P.A. & Rosenberg, J.M. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules .1. The Method. *Journal of Computational Chemistry* **13**, 1011-1021 (1992).

95. 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**, 1339-1350 (1995).

96. Roux, B. The Calculation of the Potential of Mean Force Using Computer-Simulations. *Computer Physics Communications* **91**, 275-282 (1995).

97. Song, K. et al. An Improved Reaction Coordinate for Nucleic Acid Base Flipping Studies. *Journal of Chemical Theory and Computation* **5**, 3105-3113 (2009).

98. Zharkov, D.O. & Grollman, A.P. The DNA trackwalkers: Principles of lesion search and recognition by DNA glycosylases. *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis* **577**, 24-54 (2005).

99. Huang, N., Banavali, N.K. & MacKerell, A.D. Probing the mechanism of base flipping in DNA via potential of mean force calculations. *Abstracts of Papers of the American Chemical Society* **224**, U491-U491 (2002).

100. Hegde, M.L., Hazra, T.K. & Mitra, S. Early steps in the DNA base excision/single-strand interruption repair pathway in mammalian cells. *Cell Research* **18**, 27-47 (2008).

101. Scharer, O. & Schultz, C. Closing the gap between chemistry and biology. *Chembiochem* **6**, 3-5 (2005).

102. Ferentz, A.E., Keating, T.A. & Verdine, G.L. Synthesis and Characterization of Disulfide Cross-Linked Oligonucleotides. *Journal of the American Chemical Society* **115**, 9006-9014 (1993).

103. Erlanson, D.A., Chen, L. & Verdine, G.L. DNA Methylation through a Locally Unpaired Intermediate. *Journal of the American Chemical Society* **115**, 12583-12584 (1993).

104. Harwood, E.A., Sigurdsson, S.T., Edfeldt, N.B.F., Reid, B.R. & Hopkins, P.B. Chemical synthesis and preliminary structural characterization of a nitrous acid interstrand cross-linked duplex DNA. *Journal of the American Chemical Society* **121**, 5081-5082 (1999).

105. Kobertz, W.R. & Essigmann, J.M. Solid-phase synthesis of oligonucleotides containing a site-specific psoralen derivative. *Journal of the American Chemical Society* **119**, 5960-5961 (1997).

106. Dooley, P.A. et al. Structural studies of an oligodeoxynucleotide containing a trimethylene interstrand cross-link in a 5 '-(CpG) motif: Model of a malondialdehyde cross-link. *Journal of the American Chemical Society* **123**, 1730-1739 (2001).

107. Noll, D.M., Noronha, A.M. & Miller, P.S. Synthesis and characterization of DNA duplexes containing an (NC)-C-4-ethyl-(NC)-C-4 interstrand cross-link. *Journal of the American Chemical Society* **123**, 3405-3411 (2001).

108. Wilds, C.J., Noronha, A.M., Robidoux, S. & Miller, P.S. Mispair-aligned (NT)-T-3-alkyl-(NT)-T-3 interstrand cross-linked DNA: Synthesis and characterization of duplexes with interstrand cross-links of variable lengths. *Journal of the American Chemical Society* **126**, 9257-9265 (2004).

109. Hong, I.S. & Greenberg, M.M. Efficient DNA interstrand cross-link formation from a nucleotide radical. *Journal of the American Chemical Society* **127**, 3692-3693 (2005).

110. Alzeer, J. & Scharer, O.D. A modified thymine for the synthesis of site-specific thymine-guanine DNA interstrand crosslinks. *Nucleic Acids Research* **34**, 4458-4466 (2006).

111. Wilds, C.J., Xu, F. & Noronha, A.M. Synthesis and characterization of DNA containing an N-1-2 '-deoxyinosine-ethyl-N-3-thymidine interstrand cross-link: A structural mimic of the cross-link formed by 1,3-bis-(2-chloroethyl)-1-nitrosourea. *Chemical Research in Toxicology* **21**, 686-695 (2008).

112. Muniandy, P.A., Liu, J., Majumdar, A., Liu, S.T. & Seidman, M.M. DNA interstrand crosslink repair in mammalian cells: step by step. *Critical Reviews in Biochemistry and Molecular Biology* **45**, 23-49 (2010).

113. Ho, V. & Scharer, O.D. Translesion DNA Synthesis Polymerases in DNA Interstrand Crosslink Repair. *Environmental and Molecular Mutagenesis* **51**, 552-566 (2010).

114. Hlavin, E.M., Smeaton, M.B. & Miller, P.S. Initiation of DNA Interstrand Cross-link Repair in Mammalian Cells. *Environmental and Molecular Mutagenesis* **51**, 604-624 (2010).

115. Song, K., Hornak, V., De los Santos, C., Grollman, A.P. & Simmerling, C. Molecular mechanics parameters for the FapydG DNA lesion. *Journal of Computational Chemistry* **29**, 17-23 (2008).

116. Wang, J.M., 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**, 1049-1074 (2000).

117. Cieplak, P., Cornell, W.D., Bayly, C. & Kollman, P.A. Application of the Multimolecule and Multiconformational Resp Methodology to Biopolymers - Charge Derivation for DNA, Rna, and Proteins. *Journal of Computational Chemistry* **16**, 1357-1377 (1995).

118. Frisch, M.J. et al. *Gaussian, Inc., Wallingford CT* (2004).

119. Hornak, V. et al. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712-25 (2006).

120. Case, D.A. et al. The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668-88 (2005).

121. Ryckaert, J.P., Ciccotti, G. & Berendsen, H.J.C. Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *Journal of Computational Physics* **23**, 327-341 (1977).

122. Berendsen, H.J.C., Postma, J.P.M., Vangunsteren, W.F., Dinola, A. & Haak, J.R. Molecular-Dynamics with Coupling to an External Bath. *Journal of Chemical Physics* **81**, 3684-3690 (1984).

123. Lavery, R. & Sklenar, H. Defining the Structure of Irregular Nucleic-Acids - Conventions and Principles. *Journal of Biomolecular Structure & Dynamics* **6**, 655-667 (1989).

124. Lee, S., Bowman, B.R., Ueno, Y., Wang, S. & Verdine, G.L. Synthesis and structure of duplex DNA containing the genotoxic nucleobase lesion N7-methylguanine. *Journal of the American Chemical Society* **130**, 11570-+ (2008).

125. Ojwang, J.O., Grueneberg, D.A. & Loechler, E.L. Synthesis of a Duplex Oligonucleotide Containing a Nitrogen-Mustard Interstrand DNA-DNA Cross-Link. *Cancer Research* **49**, 6529-6537 (1989).

126. Millard, J.T., Raucher, S. & Hopkins, P.B. Mechlorethamine Cross-Links Deoxyguanosine Residues at 5'-Gnc Sequences in Duplex DNA Fragments. *Journal of the American Chemical Society* **112**, 2459-2460 (1990).

127. Rink, S.M. & Hopkins, P.B. A Mechlorethamine-Induced DNA Interstrand Cross-Link Bends Duplex DNA. *Biochemistry* **34**, 1439-1445 (1995).

128. Dong, Q. et al. A structural basis for a phosphoramide mustard-induced DNA interstrand cross-link at 5'-d(GAC). *Proc Natl Acad Sci U S A* **92**, 12170-4 (1995).

129. Khullar, S., Varaprasad, C.V. & Johnson, F. Postsynthetic generation of a major acrolein adduct of 2 '-deoxyguanosine in oligomeric DNA. *Journal of Medicinal Chemistry* **42**, 947-950 (1999).

130. Hofr, C. & Brabec, V. Thermal and thermodynamic properties of duplex DNA containing site-specific interstrand cross-link of antitumor cisplatin or its clinically ineffective trans isomer. *Journal of Biological Chemistry* **276**, 9655-9661 (2001).

131. Raschle, M. et al. Mechanism of replication-coupled DNA interstrand crosslink repair. *Cell* **134**, 969-980 (2008).

132. Knipscheer, P. et al. The Fanconi Anemia Pathway Promotes Replication-Dependent DNA Interstrand Cross-Link Repair. *Science* **326**, 1698-1701 (2009).

133. Goodman, L.S., Wintrobe, M.M., Dameshek, W., Goodman, M.J. & Gilman, A. Nitrogen Mustard Therapy - Use of Methyl-Bis(Beta-Chloroethyl)Amine Hydrochloride and Tris(Beta-Chloroethyl)Amine Hydrochloride for Hodgkins Disease, Lymphosarcoma, Leukemia and Certain Allied and Miscellaneous Disorders. *Jama-Journal of the American Medical Association* **132**, 126-132 (1946).

134. Roberts, R.J. & Cheng, X. Base flipping. *Annu Rev Biochem* **67**, 181-98 (1998).

135. Klimasauskas, S., Kumar, S., Roberts, R.J. & Cheng, X.D. Hhal Methyltransferase Flips Its Target Base out of the DNA Helix. *Cell* **76**, 357-369 (1994).

136. Reinisch, K.M., Chen, L., Verdine, G.L. & Lipscomb, W.N. The Crystal-Structure of Haeiii Methyltransferase Covalently Complexed to DNA - an Extrahelical Cytosine and Rearranged Base-Pairing. *Cell* **82**, 143-153 (1995).

137. Sugahara, M. et al. Crystal structure of a repair enzyme of oxidatively damaged DNA, MutM (Fpg), from an extreme thermophile, Thermus thermophilus HB8. *Embo Journal* **19**, 3857-3869 (2000).

138. Fromme, J.C. & Verdine, G.L. Structural insights into lesion recognition and repair by the bacterial 8-oxoguanine DNA glycosylase MutM. *Nat Struct Biol* **9**, 544-52 (2002).

139. Serre, L., de Jesus, 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**, 2854-2865 (2002).

140. Leroy, J.L., Kochoyan, M., Huynhdinh, T. & Gueron, M. Characterization of Base-Pair Opening in Deoxynucleotide Duplexes Using Catalyzed Exchange of the Imino Proton. *Journal of Molecular Biology* **200**, 223-238 (1988).

141. Coman, D. & Russu, I.M. A nuclear magnetic resonance investigation of the energetics of basepair opening pathways in DNA. *Biophysical Journal* **89**, 3285-3292 (2005).

142. Giudice, E., Varnai, 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**, 1434-1443 (2003).

143. Song, K., Hornak, V., de los Santos, C., Grollman, A.P. & Simmerling, C. Molecular mechanics parameters for the FapydG DNA lesion. *J Comput Chem* **29**, 17-23 (2008).

144. Beveridge, D.L. et al. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(C(p)G) steps. *Biophysical Journal* **87**, 3799-3813 (2004).

145. Kottalam, J. & Case, D.A. Dynamics of Ligand Escape from the Heme Pocket of Myoglobin. *Journal of the American Chemical Society* **110**, 7690-7697 (1988).

146. Krueger, A., Protozanova, E. & Frank-Kamenetskii, M.D. Sequence-dependent basepair opening in DNA double helix. *Biophysical Journal* **90**, 3091-3099 (2006).

147. Serre, L., Pereira de Jesus, K., Boiteux, S., Zelwer, C. & Castaing, B. Crystal structure of the Lactococcus lactis formamidopyrimidine-DNA glycosylase bound to an abasic site analogue-containing DNA. *EMBO J* **21**, 2854-65 (2002).

148. Fromme, P. & Sayir, M.B. Detection of cracks at rivet holes using guided waves. *Ultrasonics* **40**, 199-203 (2002).

149. Coste, F. et al. Structural basis for the recognition of the FapydG lesion (2,6-diamino-4-hydroxy-5-formamidopyrimidine) by formamidopyrimidine-DNA glycosylase. *J Biol Chem* **279**, 44074-83 (2004).

150. Pereira de Jesus, K., Serre, L., Zelwer, C. & Castaing, B. Structural insights into abasic site for Fpg specific binding and catalysis: comparative high-resolution crystallographic studies of Fpg bound to various models of abasic site analogues-containing DNA. *Nucleic Acids Res* **33**, 5936-44 (2005).

151. Song, K., Kelso, C., de los Santos, C., Grollman, A.P. & Simmerling, C. Molecular Simulations reveal a common binding mode for glycosylase binding of oxidatively damaged DNA lesions. *Journal of the American Chemical Society* **129**, 14536-+ (2007).

152. Amara, P., Serre, L., Castaing, B. & Thomas, A. Insights into the DNA repair process by the formamidopyrimidine-DNA glycosylase investigated by molecular dynamics. *Protein Sci* **13**, 2009-21 (2004).

153. 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**, 16092-16105 (2004).

154. Amara, P. & Serre, L. Functional flexibility of Bacillus stearothermophilus formamidopyrimidine DNA-glycosylase. *DNA Repair (Amst)* **5**, 947-58 (2006).

155. Song, K., Hornak, V., de Los Santos, C., Grollman, A.P. & Simmerling, C. Computational analysis of the mode of binding of 8-oxoguanine to formamidopyrimidine-DNA glycosylase. *Biochemistry* **45**, 10886-94 (2006).

156. Fedorova, O.S. et al. Stopped-flow kinetic studies of the interaction between Escherichia coli Fpg protein and DNA substrates. *Biochemistry* **41**, 1520-8 (2002).

157. Koval, V.V. et al. Pre-steady-state kinetics shows differences in processing of various DNA lesions by Escherichia coli formamidopyrimidine-DNA glycosylase. *Nucleic Acids Res* **32**, 926-35 (2004).

158. Kuznetsov, N.A. et al. Kinetic conformational analysis of human 8-oxoguanine-DNA glycosylase. *J Biol Chem* **282**, 1029-38 (2007).

159. Fromme, J.C., Bruner, S.D., Yang, W., Karplus, M. & Verdine, G.L. Product-assisted catalysis in base-excision DNA repair. *Nat Struct Biol* **10**, 204-11 (2003).

160. Kuznetsov, N.A. et al. Pre-steady-state kinetic study of substrate specificity of Escherichia coli formamidopyrimidine--DNA glycosylase. *Biochemistry* **46**, 424-35 (2007).

161. Adcock, S.A. & McCammon, J.A. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem Rev* **106**, 1589-615 (2006).

162. Marks, L.B. et al. The impact of advanced technologies on treatment deviations in radiation treatment delivery. *International Journal of Radiation Oncology Biology Physics* **69**, 1579-1586 (2007).

163. Mathews, D.H. & Case, D.A. Nudged elastic band calculation of minimal energy paths for the conformational change of a GG non-canonical pair. *Journal of Molecular Biology* **357**, 1683-1693 (2006).

164. 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**, 6331-6336 (2003).

165. 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).

166. Radom, C.T., Banerjee, A. & Verdine, G.L. Structural characterization of human 8-oxoguanine DNA glycosylase variants bearing active site mutations. *J Biol Chem* **282**, 9182-94 (2007).

167. Parker, J.B. et al. Enzymatic capture of an extrahelical thymine in the search for uracil in DNA. *Nature* **449**, 433-7 (2007).

168. Parikh, S.S. et al. Base excision repair initiation revealed by crystal structures and binding kinetics of human uracil-DNA glycosylase with DNA. *EMBO J* **17**, 5214-26 (1998).

169. Chmiel, N.H., Livingston, A.L. & David, S.S. Insight into the functional consequences of inherited variants of the hMYH adenine glycosylase associated with colorectal cancer: complementation assays with hMYH variants and pre-steady-state kinetics of the corresponding mutated E.coli enzymes. *J Mol Biol* **327**, 431-43 (2003).

170. 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**, 12039-49 (2006).

171. Jiang, Y.L. & Stivers, J.T. Mutational analysis of the base-flipping mechanism of uracil DNA glycosylase. *Biochemistry* **41**, 11236-47 (2002).

172. Lavrukhin, O.V. et al. Topoisomerase II from Chlorella virus PBCV-1. Characterization of the smallest known type II topoisomerase. *J Biol Chem* **275**, 6915-21 (2000).

173. Zharkov, D.O. et al. Structural analysis of an Escherichia coli endonuclease VIII covalent reaction intermediate. *EMBO J* **21**, 789-800 (2002).

174. Burgess, S., Jaruga, P., Dodson, M.L., Dizdaroglu, M. & Lloyd, R.S. Determination of active site residues in Escherichia coli endonuclease VIII. *J Biol Chem* **277**, 2938-44 (2002).

175. Rachofsky, E.L., Osman, R. & Ross, J.B. Probing structure and dynamics of DNA with 2-aminopurine: effects of local environment on fluorescence. *Biochemistry* **40**, 946-56 (2001).

176. 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. *Proc Natl Acad Sci U S A* **97**, 13573-8 (2000).

177. Bruner, S.D., Norman, D.P., Fromme, J.C. & Verdine, G.L. Structural and mechanistic studies on repair of 8-oxoguanine in mammalian cells. *Cold Spring Harb Symp Quant Biol* **65**, 103-11 (2000).

178. Fromme, J.C., Banerjee, A., Huang, S.J. & Verdine, G.L. Structural basis for removal of adenine mispaired with 8-oxoguanine by MutY adenine DNA glycosylase. *Nature* **427**, 652-6 (2004).

179. Lindahl, T. Instability and Decay of the Primary Structure of DNA. *Nature* **362**, 709-715 (1993).

180. Kunkel, T.A. The high cost of living. American Association for Cancer Research Special Conference: endogenous sources of mutations, Fort Myers, Florida, USA, 11-15 November 1998. *Trends in Genetics* **15**, 93-4 (1999).

181. Savva, R., Mcauleyhecht, K., Brown, T. & Pearl, L. The Structural Basis of Specific Base-Excision Repair by Uracil-DNA Glycosylase. *Nature* **373**, 487-493 (1995).

182. Mol, C.D. et al. Crystal-Structure and Mutational Analysis of Human Uracil-DNA Glycosylase - Structural Basis for Specificity and Catalysis. *Cell* **80**, 869-878 (1995).

183. Parikh, S.S. et al. Uracil-DNA glycosylase-DNA substrate and product structures: Conformational strain promotes catalytic efficiency by coupled stereoelectronic effects. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 5083-5088 (2000).

184. 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).

185. Lau, A.Y., Scharer, O.D., Samson, L., Verdine, G.L. & Ellenberger, E. Crystal structure of a human alkylbase-DNA repair enzyme complexed to DNA: Mechanisms for nucleotide flipping and base excision. *Cell* **95**, 249-258 (1998).

186. 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**, 13573-13578 (2000).

187. Chetsanga, C.J. & Lindahl, T. Release of 7-Methylguanine Residues Whose Imidazole Rings Have Been Opened from Damaged DNA by a DNA Glycosylase from Escherichia-Coli. *Nucleic Acids Research* **6**, 3673-3684 (1979).

188. Chetsanga, C.J., Lozon, M., Makaroff, C. & Savage, L. Purification and Characterization of Escherichia-Coli Formamidopyrimidine-DNA Glycosylase That Excises Damaged 7-Methylguanine from Deoxyribonucleic-Acid. *Biochemistry* **20**, 5201-5207 (1981).

189. Fromme, J.C. & Verdine, G.L. Structural insights into lesion recognition and repair by the bacterial 8-oxoguanine DNA glycosylase MutM. *Nature Structural Biology* **9**, 544-552 (2002).

190. Zharkov, D.O., Shoham, G. & Grollman, A.P. Structural characterization of the Fpg family of DNA glycosylases. *DNA Repair* **2**, 839-862 (2003).

191. Shibutani, S., Takeshita, M. & Grollman, A.P. Insertion of Specific Bases during DNA-Synthesis Past the Oxidation-Damaged Base 8-Oxodg. *Nature* **349**, 431-434 (1991).

192. van Loon, B., Markkanen, E. & Hubscher, U. Oxygen as a friend and enemy: How to combat the mutational potential of 8-oxo-guanine. *DNA Repair* **9**, 604-616 (2010).

193. Kasai, H. & Nishimura, S. Hydroxylation of Deoxyguanosine at the C-8 Position by Ascorbic-Acid and Other Reducing Agents. *Nucleic Acids Research* **12**, 2137-2145 (1984).

194. De Bont, R. & van Larebeke, N. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19**, 169-185 (2004).

195. Brieba, L.G. et al. Structural basis for the dual coding potential of 8-oxoguanosine by a high-fidelity DNA polymerase. *Embo Journal* **23**, 3452-3461 (2004).

196. Nash, H.M. et al. Cloning of a yeast 8-oxoguanine DNA glycosylase reveals the existence of a base-excision DNA-repair protein superfamily. *Current Biology* **6**, 968-980 (1996).

197. Nash, H.M., Lu, R.Z., Lane, W.S. & Verdine, G.L. The critical active-site amine of the human 8-oxoguanine DNA glycosylase, hOgg1: direct identification, ablation and chemical reconstitution. *Chemistry & Biology* **4**, 693-702 (1997).

198. Dodson, M.L., Michaels, M.L. & Lloyd, R.S. Unified Catalytic Mechanism for DNA Glycosylases. *Journal of Biological Chemistry* **269**, 32709-32712 (1994).

199. Loncharich, R.J., Brooks, B.R. & Pastor, R.W. Langevin Dynamics of Peptides - the Frictional Dependence of Isomerization Rates of N-Acetylalanyl-N'-Methylamide. *Biopolymers* **32**, 523-535 (1992).

200. Song, K. et al. An Improved Reaction Coordinate for Nucleic Acid Base Flipping Studies *J. Chem. Theo. Comput.* **5**, 3105-3113 (2009).

201. Friedman, J.I. & Stivers, J.T. Detection of Damaged DNA Bases by DNA Glycosylase Enzymes. *Biochemistry* **49**, 4957-4967 (2010).

202. Ravishanker, G., Swaminathan, S., Beveridge, D.L., Lavery, R. & Sklenar, H. Conformational and Helicoidal Analysis of 30 Ps of Molecular-Dynamics on the D(Cgcgaattcgcg) Double Helix - Curves, Dials and Windows. *Journal of Biomolecular Structure & Dynamics* **6**, 669-699 (1989).

203. Guainazzi, A., Campbell, A.J., Angelov, T., Simmerling, C. & Scharer, O.D. Synthesis and Molecular Modeling of a Nitrogen Mustard DNA Interstrand Crosslink. *Chemistry-a European Journal* **16**, 12100-12103 (2010).

Appendix 1- Z-matrix Coordinates and Charges for Analog Nitrogen Mustard (AN1)

```
   0    0    2

This is a remark line
molecule.res
AN1    INT  0
CORRECT     OMIT DU   BEG
 -1.22187
   1  DUMM  DU    M    0  -1  -2   0.000      .0        .0      .00000
   2  DUMM  DU    M    1   0  -1   1.449      .0        .0      .00000
   3  DUMM  DU    M    2   1   0   1.522   111.1        .0      .00000
   4  O1P   O2    M    3   2   1   1.540   111.208   180.000  -0.77610
   5  P     P     M    4   3   2   1.549   106.189    12.458   1.16590
   7  O2P   O2    E    5   4   3   1.760   109.501   -84.533  -0.77610
   8  O5'   OS    M    5   4   3   1.761   109.494    35.493  -0.49540
   9  C5'   CT    M    9   5   4   1.429   109.301    60.124  -0.00690
  10  H5'1  H1    E   10   9   5   1.070   109.503    59.455   0.07540
  11  H5'2  H1    E   10   9   5   1.071   109.421   -60.535   0.07540
  12  C4'   CT    M   10   9   5   1.541   109.615   179.472   0.16290
  13  H4'   H1    E   13  10   9   1.070   108.837   178.793   0.11760
  14  C3'   CT    3   13  10   9   1.541   113.559    58.210   0.07130
  15  C2'   CT    B   15  13  10   1.484   100.995   -88.039  -0.08540
  16  H2'1  HC    E   16  15  13   1.070   112.866    67.468   0.07180
  17  H2'2  HC    E   16  15  13   1.070   114.162  -169.656   0.07180
  18  H3'   H1    E   15  13  10   1.070   113.760    37.716   0.09850
  19  O3'   OS    S   15  13  10   1.429   109.891   159.280  -0.52320
  21  O4'   OS    M   13  10   9   1.473   106.778   -56.964  -0.36910
  22  C1'   CT    M   23  13  10   1.447   104.245   128.005   0.03580
  23  H1'   H2    E   24  23  13   1.071   110.955    78.399   0.17460
  24  N9    N*    M   24  23  13   1.470   112.476  -160.211  -0.07266
  25  C4    CB    S   26  24  23   1.340   126.662  -151.505   0.28181
  26  C5    CB    S   27  26  24   1.341   109.045   179.921  -0.13112
  27  C6    C     B   28  27  26   1.379   120.531  -179.914   0.49587
  28  O6    O     E   29  28  27   1.258   122.199  -179.851  -0.55588
  29  N1    NA    B   29  28  27   1.467   115.560     0.161  -0.44878
  30  H1    H     E   31  29  28   1.000   120.299   179.878   0.31997
  31  C2    CA    B   31  29  28   1.357   119.416    -0.115   0.67750
  32  N2    N2    B   33  31  29   1.337   119.154   179.837  -0.91284
  33  H21   H     E   34  33  31   1.000   119.974   179.954   0.39030
  34  H22   H     E   34  33  31   1.000   120.003    -0.076   0.39030
  35  N3    NC    E   33  31  29   1.315   121.520    -0.157  -0.60485
  36  C8    CK    M   26  24  23   1.467   126.270    28.419  -0.26430
  37  H8    H5    E   38  26  24   1.070   126.588    -0.004   0.20144
  38  C7    C*    M   38  26  24   1.352   106.747  -179.954   0.09219
  39  C15   CL    M   40  38  26   1.539   126.328  -179.960   0.07419
  40  H15   HL    E   41  40  38   1.069   109.577  -149.289   0.06480
  41  H25   HL    E   41  40  38   1.070   109.523    90.601   0.06480
  42  N14   NL    M   41  40  38   1.470   109.210   -29.199  -0.33188
  43  H14   HM    E   44  41  40   1.001   109.456   -60.223   0.23560
  44  N13   NL    M   44  41  40   1.400   109.750   179.843  -0.60474
```

```
  45  H13   HM    E   46  44  41   0.999   109.562    60.813   0.32761

LOOP
  C1' C2'
  C4  C5
  C4  N9

IMPROPER
  C8   C4   N9   C1'
  C5   N1   C6   O6
  C6   C2   N1   H1
  C2   H21  N2   H22
  C7   N9   C8   H8
  N1   N3   C2   N2
  C5   C8   C7   C15

DONE
STOP
```

Appendix 2- Z-matrix Coordinates and Charges for Analog Nitrogen Mustard (AN2)

```
   0    0    2

This is a remark line
molecule.res
AN2    INT  0
CORRECT     OMIT DU   BEG
 -0.77815
   1  DUMM  DU    M    0  -1  -2   0.000      .0        .0      .00000
   2  DUMM  DU    M    1   0  -1   1.449      .0        .0      .00000
   3  DUMM  DU    M    2   1   0   1.522   111.1        .0      .00000
  46  C12   CL    M   44  42  39   1.469   109.215  -179.064   0.19529
  47  H12   HL    E   46  44  42   1.070   109.446    60.273   0.00447
  48  H32   HL    E   46  44  42   1.071   109.551  -179.773   0.00447
  49  C11   CL    M   46  44  42   1.538   109.447   -59.687   0.00639
  50  H11   HL    E   49  46  44   1.070   109.535   -60.382   0.03174
  51  H31   HL    E   49  46  44   1.070   109.458    59.615   0.03174
  52  C7    C*    M   49  46  44   1.539   109.425   179.522   0.09219
  53  C8    CK    S   52  49  46   1.353   126.457   -89.681  -0.26430
  54  H8    H5    E   53  52  49   1.070   126.673     0.212   0.20144
  55  C5    CB    M   52  49  46   1.402   126.603    90.379  -0.13112
  56  C6    C     M   55  52  49   1.379   129.152    -0.207   0.49587
  57  O6    O     E   56  55  52   1.259   122.237    -0.011  -0.55588
  58  N1    NA    M   56  55  52   1.467   115.463   179.995  -0.44878
  59  H1    H     E   58  56  55   1.000   120.279  -179.989   0.31997
  60  C2    CA    M   58  56  55   1.356   119.429     0.009   0.67750
  61  N2    N2    B   60  58  56   1.338   119.140  -179.995  -0.91284
  62  H21   H     E   61  60  58   1.000   120.058   179.968   0.39030
  63  H22   H     E   61  60  58   1.000   119.906    -0.049   0.39030
  64  N3    NC    M   60  58  56   1.316   121.526     0.023  -0.60485
  65  C4    CB    M   64  60  58   1.341   120.148    -0.017   0.28181
  66  N9    N*    M   65  64  60   1.341   128.089   179.948  -0.07266
  67  C1'   CT    M   66  65  64   1.470   126.719     0.147   0.03580
  68  H1'   H2    E   67  66  65   1.070   108.850    30.460   0.17460
  69  C2'   CT    3   67  66  65   1.523   113.403   150.540  -0.08540
  70  H2'1  HC    E   69  67  66   1.071   111.987    53.756   0.07180
  71  H2'2  HC    E   69  67  66   1.069   114.101   -68.786   0.07180
  72  C3'   CT    B   69  67  66   1.501    96.865   169.778   0.07130
  73  H3'   H1    E   72  69  67   1.070   114.979  -174.467   0.09850
  74  O3'   OS    S   72  69  67   1.430   111.073    63.251  -0.52320
  76  O4'   OS    M   67  66  65   1.471   113.025   -90.169  -0.36910
  77  C4'   CT    M   76  67  66   1.469   103.982  -141.824   0.16290
  78  H4'   H1    E   77  76  67   1.070   112.811  -141.359   0.11760
  79  C5'   CT    M   77  76  67   1.540   108.572    98.101  -0.00690
  80  H5'1  H1    E   79  77  76   1.069   109.533    63.512   0.07540
  81  H5'2  H1    E   79  77  76   1.070   109.449  -176.471   0.07540
  82  O5'   OS    M   79  77  76   1.429   109.534   -56.415  -0.49540
  83  P     P     M   82  79  77   1.760   109.348    59.968   1.16590
  85  O1P   O2    E   83  82  79   1.760   109.499  -179.978  -0.77610
  86  O2P   O2    M   83  82  79   1.549   109.517    59.969  -0.77610
```

```
LOOP
  C4  N3
  C4  N9
  C3' C4'

IMPROPER
  C8   C4   N9   C1'
  C5   N1   C6   O6
  C6   C2   N1   H1
  C2   H21  N2   H22
  C7   N9   C8   H8
  N1   N3   C2   N2
  C5   C8   C7   C11

DONE
STOP
```

Appendix 3- Force field parameters for Analog Nitrogen Mustard

```
remark goes here
MASS
CL 12.010       0.878             GAFF c3
HL  1.008       0.135             GAFF h1
NL 14.010       0.530             GAFF n3
HM  1.008       0.161             GAFF hn

BOND
C*-CK   546.0    1.352      C*-CW TRP CG-CD1
C*-CL   337.3    1.499      GAFF cc-CL or cd-CL
CL-HL   335.9    1.093      GAFF   c3-h1
CL-CL   303.1    1.535      GAFF   c3-c3
NL-HM   394.1    1.018      GAFF   hn-n3
NL-NL   383.6    1.454      GAFF   n3-n3
CL-NL   320.6    1.470      GAFF   c3-n3

ANGLE
CB-C*-CL   70.0      128.60   CB-C*-CT AA trp from ff99SB
CK-C*-CL   70.0      125.00   CT-C*-CW TRP CB-CG-CD1
C*-CL-HL   50.0      109.50   AA trp
C*-CL-CL   64.7      108.10   GAFF CL-CL-cc
HL-CL-HL   39.2      109.55   GAFF h1-c3-h1
HL-CL-CL   46.4      110.07   GAFF h1-c3-c3
CL-CL-NL   66.2      110.38   GAFF c3-c3-n3
CL-NL-HM   47.1      109.92   GAFF c3-n3-hn
CL-NL-NL   66.8      108.15   GAFF c3-n3-n3
HM-NL-NL   50.2      103.98   GAFF hn-n3-n3
CL-C*-CK   70.0      125.00   CT-C*-CW TRP CB-CG-CD1
CB-C*-CK   63.0      106.40   CB-C*-CW TRP CD2-CG-CD1
C*-CB-C    63.0      132.00   CA-CB-C* TRP CE3-CD2-CG then -2.9
C*-CB-CB   63.0      108.80   C*-CB-CN TRP CG-CD2-CE2
C*-CK-H5   50.0      120.00   C*-CW-H4 TRP CG-CD1-HD1
C*-CK-N*   70.0      113.20   C*-CW-NA TRP CG-CD1-NE1 then +4.5
HL-CL-NL   49.4      109.92   GAFF h1-c3-n3
C*-CL-NL   66.5      111.47   GAFF c2-c3-n3

DIHE
X-C*-CK-X     4   26.100      180.0           2.        X-C*-CW-X TRP X-CG-CD1-X
X-C*-CL-X     6    0.000       0.000        3.000      GAFF X-cd-CL-X and ff99SB X-C*-CT-X
X-CL-CL-X     9    1.400       0.000        3.000      GAFF X-c3-c3-X
X-CL-NL-X     6    1.800       0.000        3.000      GAFF X-c3-n3-X
X-NL-NL-X     4    9.000       0.000        2.000      GAFF X-n3-n3-X


IMPROPER
CB-CK-C*-CL  1.1     180.     2.                       CW-CB-C*-CT TRP

NONBON
CL         1.9080  0.1094          GAFF  c3
HL         1.3870  0.0157          GAFF  h1
NL         1.8240  0.1700          GAFF  n3
HM          0.6000  0.0157          GAFF  hn
```

Appendix 4- Z-matrix Coordinates and Charges for Nitrogen Mustard (NX1)

```
   0    0    2

This is a remark line
NX1.res
NX1    INT  0
CORRECT     OMIT DU   BEG
  0.67322
   1  DUMM  DU    M    0  -1  -2     0.000      .0        .0      .00000
   2  DUMM  DU    M    1   0  -1     1.449      .0        .0      .00000
   3  DUMM  DU    M    2   1   0     1.522   111.1        .0      .00000
   4  O1P   O2    M    3   2   1     1.540   111.208   180.000  -0.77610
   5    P   P     M    4   3   2     1.759    68.388   -99.111   1.16590
   7  O2P   O2    E    5   4   3     1.760   109.478    26.164  -0.77610
   8  O5'   OS    M    5   4   3     1.760   109.478   -93.835  -0.49540
   9  C5'   CI    M    8   5   4     1.417   109.599   161.498  -0.00690
  10  H5'1  H1    E    9   8   5     1.084   109.858    85.825   0.07540
  11  H5'2  H1    E    9   8   5     1.085   110.884   -33.752   0.07540
  12  C4'   CT    M    9   8   5     1.514   109.498  -154.224   0.16290
  13  H4'   H1    E   12   9   8     1.080   109.227  -173.565   0.11760
  14  C3'   CT    3   12   9   8     1.535   116.658    62.362   0.07130
  15  C2'   CT    B   14  12   9     1.530   103.002   -95.409  -0.08540
  16  H2'1  HC    E   15  14  12     1.082   110.464    83.639   0.07180
  17  H2'2  HC    E   15  14  12     1.081   112.763  -153.524   0.07180
  18  H3'   H1    E   14  12   9     1.083   112.041    25.608   0.09850
  19  O3'   OS    S   14  12   9     1.396   106.230   147.952  -0.52320
  21  O4'   OS    M   12   9   8     1.433   108.713   -56.183  -0.36910
  22  C1'   CT    M   21  12   9     1.377   111.750   119.926   0.03580
  23  H1'   H2    E   22  21  12     1.078   111.809   108.378   0.17460
  24   N9   N*    M   22  21  12     1.494   105.668  -137.790   0.03702
  25   C4   CB    S   24  22  21     1.372   127.285  -143.658   0.27346
  26   C5   CB    S   25  24  22     1.369   106.749   175.976  -0.11250
  27   C6   C     B   26  25  24     1.417   119.312  -177.079   0.58299
  28   O6   O     E   27  26  25     1.215   128.341   175.892  -0.53978
  29   N1   NA    B   27  26  25     1.383   110.952    -3.368  -0.51506
  30   H1   H     E   29  27  26     1.001   114.824  -178.733   0.40801
  31   C2   CA    B   29  27  26     1.373   125.567     2.206   0.89752
  32   N2   N2    B   31  29  27     1.320   117.761   179.705  -0.90523
  33  H21   H     E   32  31  29     0.995   123.782     1.150   0.45823
  34  H22   H     E   32  31  29     0.999   118.253  -179.950   0.45823
  35   N3   NC    E   31  29  27     1.310   122.313     0.078  -0.62683
  36   C8   CK    M   24  22  21     1.325   124.417    32.114   0.01844
  37   H8   H5    E   36  24  22     1.072   122.649     5.706   0.25853
  38   N7   N*    M   36  24  22     1.315   111.210  -175.828  -0.07929
  39  C15   CT    M   38  36  24     1.463   124.696  -173.809   0.14249
  40  H15   HC    E   39  38  36     1.081   108.298   117.225   0.11624
  41  H25   HC    E   39  38  36     1.079   106.416     2.250   0.11624
  42  C14   CH    M   39  38  36     1.542   114.340  -118.075   0.01675
  43  H14   HP    E   42  39  38     1.080   109.795   134.381   0.11627
  44  H24   HP    E   42  39  38     1.080   111.227    14.221   0.11627
```

```
  45  N13   N3    M   42  39  38     1.493   111.903  -106.891  -0.38564
  46  H13   H     E   45  42  39     1.021   109.753   -62.867   0.36603
  47  H23   H     E   45  42  39     1.022   107.281    54.897   0.36603


LOOP
 C1' C2'
 C4  C5
 C4  N9

IMPROPER
 C8   C4   N9   C1'
 C5   N1   C6   O6
 C6   C2   N1   H1
 C2   H21  N2   H22
 N7   N9   C8   H8
 N1   N3   C2   N2
 C5   C8   N7   C15

DONE
STOP
```

Appendix 5- Z-matrix Coordinates and Charges for Nitrogen Mustard (NX2)

```
   0    0    2

This is a remark line
NX2.res
NX2    INT  0
CORRECT     OMIT DU   BEG
  0.3268
   1  DUMM  DU    M    0  -1  -2     0.000      .0        .0      .00000
   2  DUMM  DU    M    1   0  -1     1.449      .0        .0      .00000
   3  DUMM  DU    M    2   1   0     1.522   111.1        .0      .00000
  48  C12   CH    M   45  42  39     1.492   114.087   176.648   0.01675
  49  H12   HP    E   48  45  42     1.080   107.108   -62.888   0.11627
  50  H32   HP    E   48  45  42     1.080   108.166    53.780   0.11627
  51  C11   CT    M   48  45  42     1.542   111.957   176.698   0.14249
  52  H11   HC    E   51  48  45     1.082   111.224    16.065   0.11624
  53  H31   HC    E   51  48  45     1.079   109.018   134.120   0.11624
  54  N7    N*    M   51  48  45     1.464   114.370  -106.898  -0.07929
  55  C8    CK    S   54  51  48     1.314   124.747  -118.062   0.01844
  56  H8    H5    E   55  54  51     1.072   126.116     4.491   0.25853
  57  C5    CB    M   54  51  48     1.403   128.177    70.152  -0.11250
  58  C6    C     M   57  54  51     1.416   133.445   -10.244   0.58299
  59  O6    O     E   58  57  54     1.215   128.398    -0.186  -0.53978
  60  N1    NA    M   58  57  54     1.383   110.970  -179.491  -0.51506
  61  H1    H     E   60  58  57     1.001   114.785  -178.826   0.40801
  62  C2    CA    M   60  58  57     1.373   125.535     2.270   0.89752
  63  N2    N2    B   62  60  58     1.320   117.730   179.696  -0.90523
  64  H21   H     E   63  62  60     0.996   123.834     1.412   0.45823
  65  H22   H     E   63  62  60     0.998   118.298  -179.954   0.45823
  66  N3    NC    M   62  60  58     1.309   122.359     0.003  -0.62683
  67  C4    CB    M   66  62  60     1.323   114.053    -1.079   0.27346
  68  N9    N*    M   67  66  62     1.371   125.582   179.485   0.03702
  69  C1'   CT    M   68  67  66     1.495   127.337    -3.789   0.03580
  70  H1'   H2    E   69  68  67     1.079   105.044   -25.426   0.17460
  71  C2'   CT    3   69  68  67     1.526   113.417    98.902  -0.08540
  72  H2'1  HC    E   71  69  68     1.083   110.401    30.817   0.07180
  73  H2'2  HC    E   71  69  68     1.081   112.620   -91.864   0.07180
  74  C3'   CT    B   71  69  68     1.530   100.875   147.622   0.07130
  75  H3'   H1    E   74  71  69     1.082   112.373  -154.092   0.09850
  76  O3'   OS    S   74  71  69     1.397   110.878    80.065  -0.52320
  78  O4'   OS    M   69  68  67     1.376   105.688  -143.829  -0.36910
  79  C4'   CT    M   78  69  68     1.432   111.748  -137.876   0.16290
  80  H4'   H1    E   79  78  69     1.080   107.849  -121.826   0.11760
  81  C5'   CI    M   79  78  69     1.514   108.729   119.930  -0.00690
  82  H5'1  H1    E   81  79  78     1.085   109.221    64.200   0.07540
  83  H5'2  H1    E   81  79  78     1.083   109.170  -177.704   0.07540
  84  O5'   OS    M   81  79  78     1.417   109.469   -56.102  -0.49540
  85   P    P     M   84  81  79     1.760   109.570  -154.230   1.16590
  87  O2P   O2    E   85  84  81     1.760   109.487    41.427  -0.77610
  88  O1P   O2    M   85  84  81     1.760   109.492   161.449  -0.77610
```

```
LOOP
 C4  N3
 C4  N9
 C3' C4'

IMPROPER
 C8   C4   N9   C1'
 C5   N1   C6   O6
 C6   C2   N1   H1
 C2   H21  N2   H22
 N7   N9   C8   H8
 N1   N3   C2   N2
 C5   C8   N7   C11

DONE
STOP
```

Appendix 6- Force field parameters for Nitrogen Mustard

```
remark goes here
MASS
CH 12.010       0.878             same as CT

BOND
CH-HP  337.30   1.092              same as c3-hc
CH-N3  293.60   1.499              same as c3-n4
CH-CT  303.10   1.535              same as c3-c3

ANGLE
N3-CH-CT   66.0      108.93         same as c3-c3-n4
CH-N3-CT   62.8      110.64         same as c3-n4-c3
CH-CT-HC   46.4      110.05         same as c3-c3-hc
HP-CH-CT   46.4      110.05         same as c3-c3-hc
HP-CH-HP   39.4      108.35         same as hc-c3-hc
HP-CH-N3   49.0      107.90         same as hc-c3-n4
CH-N3-H    46.2      110.11         same as c3-n4-hn
CH-CT-N*   50.0      109.50         same as CT-CT-N*
CH-N3-CH   50.0      109.50         same as CT-N3-CT
N*-CK-N*   70.0      116.80         calculated angle (force from N*-CK-NB)
N*-CB-C    70.0      134.60         calculated angle (force from C-CB-NB)
HC-CT-N*   50.0      109.50         same as same as CT-CT-N*

DIHE
X -CH-CT-X    9    1.400        0.000          3.000   same as X -c3-c3-X
X -N3-CH-X    9    1.400        0.000          3.000   same as X -c3-n4-X

IMPROPER

NONBON
CH          1.9080  0.1094          same as CT
```

Appendix 7- Z-matrix Coordinates and Charges for Etheno-deoxyguanine (EdG)

```
   0    0    2

This is a remark line
molecule.res
EDG    INT  0
CORRECT     OMIT DU   BEG
 -0.9992
   1  DUMM  DU    M    0  -1  -2     0.000      .0        .0      .00000
   2  DUMM  DU    M    1   0  -1     1.449      .0        .0      .00000
   3  DUMM  DU    M    2   1   0     1.522   111.1        .0      .00000
   4  O1P   O2    M    3   2   1     1.540   111.208   180.000  -0.77610
   5  P     P     M    4   3   2     1.760   109.483   -77.363   1.16590
   6  O2P   O2    E    5   4   3     1.760   109.454    59.993  -0.77610
   7  O5'   OS    M    5   4   3     1.040   109.483   180.000  -0.49540
   8  C5'   CI    M    7   5   4     1.430   180.000   -12.637  -0.00690
   9  H5'1  H1    E    8   7   5     1.089   109.633   -90.000   0.07540
  10  H5'2  H1    E    8   7   5     1.090   109.061   151.097   0.07540
  11  C4'   CT    M    8   7   5     1.533   112.192    31.282   0.16290
  12  H4'   H1    E   11   8   7     1.089   105.961   166.203   0.11760
  13  C3'   CT    3   11   8   7     1.541   111.780    48.504   0.07130
  14  C2'   CT    B   13  11   8     1.533   102.326  -122.278  -0.08540
  15  H2'1  HC    E   14  13  11     1.090   109.002    94.325   0.07180
  16  H2'2  HC    E   14  13  11     1.090   110.971  -144.483   0.07180
  17  H3'   H1    E   13  11   8     1.090   110.141    -2.274   0.09850
  18  O3'   OS    E   13  11   8     1.435   111.119   119.002  -0.52320
  19  O4'   OS    M   11   8   7     1.448   112.900   -72.237  -0.36910
  20  C1'   CT    M   19  11   8     1.440   108.951   146.151   0.03580
  21  H1'   H2    E   20  19  11     1.090   112.007    79.154   0.17460
  22  N9    N*    M   20  19  11     1.489   111.062  -163.856  -0.04770
  23  C4    CB    M   22  20  19     1.375   122.834   -66.403   0.28530
  24  N3    NC    M   23  22  20     1.351   124.729   -10.563  -0.58680
  25  C2    CB    M   24  23  22     1.356   110.840  -179.974   0.28940
  26  N2    NN    M   25  24  23     1.353   118.011   179.979  -0.17190
  27  H21   H     E   26  25  24     1.090   128.507     0.102   0.31200
  28  C10   CW    M   26  25  24     1.374   100.133  -179.958  -0.25370
  29  H10   H4    E   28  26  25     1.091   123.611  -179.908   0.24020
  30  C11   CV    M   28  26  25     1.376   112.803     0.053  -0.18910
  31  H11   H4    E   30  28  26     1.091   126.199  -179.959   0.22120
  32  N1    N*    M   30  28  26     1.373   106.908    -0.107   0.06350
  33  C6    C     M   32  30  28     1.424   132.127  -179.987   0.44920
  34  O6    O     E   33  32  30     1.248   123.953     0.013  -0.52580
  35  C5    CB    M   33  32  30     1.425   110.252  -179.848   0.07880
  36  N7    NB    M   35  33  32     1.387   129.177   179.981  -0.49600
  37  C8    CK    M   36  35  33     1.303   104.418  -180.000   0.08170
  38  H8    H5    E   37  36  35     1.089   120.003   179.253   0.16170


IMPROPER
 C8   C4   N9   C1'
```

```
 C5   N1   C6   O6
 C6   C2   N1   C11
 C2   H21  N2   C10
 N7   N9   C8   H8
 N1   N3   C2   N2
 C11  N2   C10  H10
 N1   C10  C11  H11

LOOP CLOSING EXPLICIT
 C1'  C2'
 C4   C5
 C4   N9
 C10  C11

DONE
STOP
```

Appendix 8- Force field parameters for Etheno-deoxyguanine (EdG)

```
remark goes here
MASS
NN 14.010       0.530             same as NA

BOND
CV-N*  410.00   1.394      same as CV-NB
CB-NN  436.00   1.374      same as CB-N*
NN-CW  427.00   1.381      same as NA-CW
CV-CW  546.00   1.352      same as C*-CW
NN-H   434.00   1.010      same as NA-H


ANGLE
CB-CB-C    70.0     119.60    CALCULATED
CB-C-N*    70.0     110.30    CALCULATED
C-N*-CB    70.0     124.90    CALCULATED
CB-NC-CB   70.0     111.40    CALCULATED
NC-CB-CB   70.0     127.60    CALCULATED
CB-N*-CV   70.0     103.20    CALCULATED
N*-CV-CW   70.0     106.90    CALCULATED
CV-CW-NN   70.0     114.60    CALCULATED
CW-NN-CB   70.0      99.10    CALCULATED
NN-CB-N*   70.0     116.20    CALCULATED
C-N*-CV    70.0     131.80    CALCULATED
N*-CV-H4   50.0     126.90    CALCULATED
H4-CV-CW   50.0     126.20    CALCULATED
CV-CW-H4   50.0     125.40    CALCULATED
CW-NN-H    50.0     131.90    CALCULATED
CB-NN-H    50.0     129.00    CALCULATED
NN-CB-NC   70.0     117.60    CALCULATED
NN-CW-H4   50.0     120.00    CALCULATED

DIHE
X-N*-CV-X     4    6.800      180.000         2.000     same as X -N*-CK-X
X-CV-CW-X     4   20.600      180.000         2.000     same as X -CC-CV-X
X-CW-NN-X     4    6.000      180.000         2.000     same as X -CW-NA-X
X-NN-CB-X     4    6.000      180.000         2.000     same as X -CW-NA-X

IMPROPER
C-CB-N*-CV    1.1       180.0       2.0
CB-H-NN-CW    1.1       180.0       2.0


NONBON
NN         1.8240  0.1700           same as N
```