Stony Brook University

Applying Computational Methods in the Study of Biomolecular Systems:

The Recognition Mechanism of DNA Repair Enzyme Fpg

A Dissertation Presented

by

Kun Song

to

The Graduate School

in Partial fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Chemistry

Stony Brook University

May 2007

Stony Brook University

The Graduate School

Kun Song,

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree,

hereby recommend acceptance of this dissertation.

Carlos L. Simmerling

Associated Professor, Department of Chemistry

Orlando Schärer

Associated Professor, Department of Chemistry

Fernando O. Raineri

Adjunct Assistant Professor, Department of Chemistry

Arthur P. Grollman

Distinguished Professor, Pharmacological Sciences

Evelyn G. Glick Professor of Experimental Medicine

Director: Laboratory for Chemical Biology

Stony Brook University

This dissertation is accepted by the Graduate School

Lawrence Martin

Dean of the Graduate School

Abstract of the Dissertation

Applying Computational Methods in the Study of Biomolecular Systems:

The Recognition Mechanism of DNA Repair Enzyme Fpg

by

Kun Song

Doctor of Philosophy

in

Chemistry

Stony Brook University

2007

8-oxo-guanine (8OG) is one of the most prevalent forms of oxidative DNA damage. Failure to repair 8OG will lead to cancer and many age-related diseases. For studying the pathophysiology of these diseases, it is essential to understand the repair mechanism of 8OG in healthy cells. In bacteria, 8OG is excised by formamidopyrimidine glycosylase (Fpg) as the initial step in base excision repair. To efficiently excise this lesion, Fpg must discriminate between 8OG and an excess of guanine in duplex DNA. We applied computational methods studying the structural basis underlying this high degree of

selectivity.

In first study free energy calculation methods and point mutation studies have been performed on the comparison of the two binding mode of 8OG. Two different binding modes of 8OG in Fpg/DNA complex have been shown in different structural studies. Our all-atom simulations are consistent with both structures. The *syn* conformation observed in the crystallographic structure of Fpg obtained from *B. stearothermophilus* is stabilized through interaction with E77, a non-conserved residue. Replacement of E77 by Ser, creating the Fpg sequence found in *E. coli* and other bacteria, results in preferred binding of 8OG in the *anti* conformation.

FapydG is another common oxidative DNA lesion involving opening of the imidazole ring. It has similar structure and shares the same precursor with 8OGG and can be excised by the same enzymes as 8OG. We examined the current force field parameters for FapydG and found that the energy barrier of the rotational bond C5-N7 is overestimated. New parameters were calculated and simulations with them can well reproduce the x-ray structures.

DNA sliding and base flipping are two essential motions in DNA lesion searching and recognition. We used targeted MD, umbrella sampling, and long MD simulations (1.6 ms in total) to simulate these two processes. We observed the sliding motions and base pair breaking in our long MD simulations. In the targeted MD simulations, after the forced conformational changes occurred, we observed that the structures of several key residues changed from the original conformations to the new ones, which reproduced the

structural differences between x-ray structures at different stages.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

8OG	8-oxo-Guanine
ANM	Anisotropic Network Model
B. st.	*Bacillus stearothermophilus*
BER	Base excision repair
CD	circular dichroism
*c*FapydG	carbo-FapydG
COM	Center of mass
DMP	Dimethylphosphate
DNA	Deoxyribonucleic acid
E. coli	*Escherichia coli*
EGB	Generalized Born energy
FapydG	2,6-diamino-4-hydroxy-5-formamidopyrimidine
Fpg	Formamidopyrimidine-DNA Glycosylase
GB	Generalized Born
HF	Hartree-Fock
hOGG1	Human 8-oxo-Guanine DNA Glycosylase
L.l.	*Lactococcus lactis*
MD	Molecular dynamics
MM	Molecular mechanics
MMGBSA	Molecular mechanics- Generalized Born Surface Area

MMPBSA	Molecular mechanics- Poisson Bolzmann Surface Area
MMR	Mismatch repair
MP2	Møller-Plesset perturbation theory of the second order
NER	Nucleotide excision repair
NMA	Normal mode analysis
NMR	Nuclear magnetic resonance
PB	Poisson Bolzmann
PCA	Principal component analysis
PDB	Protein data bank
PME	Particle Mesh Ewald
PMF	Potential of mean force
REMD	Replica exchange molecular dynamics
RESP	Restrained electrostatic potential
RMSD	Root mean square deviation
SASA	Solvent accessible surface area
TIP3P	Transferable intermolecular potential 3 points
TMD	Targeted molecular dynamics
vdw	van der Waals
WHAM	Weighted histogram analysis method

Acknowledgements

A journey is easier when you travel together. Interdependence is certainly more valuable than independence. This thesis is the result of about five and half years of work whereby I have been accompanied and supported by many people. It is a pleasant aspect that I have now the opportunity to express my gratitude for all of them.

The first person I would like to thank is my advisor Dr. Carlos Simmerling. I have been in his lab since 2001 when I started my PhD program and actively since 2002. During these years I have known Carlos as a sympathetic and principle-centered person. His overly enthusiasm and integral view on research and his mission for providing 'only high-quality work and not less', has made a deep impression on me. I owe him lots of gratitude for having me shown this way of research. He could not even realize how much I have learned from him. Besides of being an excellent supervisor, Carlos is an excellent mentor and a good friend to me. I am really glad that I have come to get know Dr. Carlos Simmerling in my life.

I would like to thank our experimental collaborators. In my first years of graduate life in Simmerling Lab I was working on the protein folding project and would like to thank Dr. Neils Anderson who was our collaborator for his helpful discussions and experimental support. Later, I was working on the DNA repair project and came to know Dr. Arthur Grollman and Dr. Carlos de los Santos who kept an eye on the progress of my work and always were available when I needed their advises. I would like to thank them

for their time as we met regularly for 2 years every other week. I'd like to take this opportunity to thank Dr. Arthur Grollman who also served as the external member in my thesis committee.

I would also like to thank the other members of my PhD committee who monitored my work and took effort in reading and providing me with valuable comments on earlier versions of this thesis: Dr. Orlando Schärer, and Dr. Fernando Raineri. I would also like to thank Dr. Orlando Schärer for letting me attending the weekly journal club hosted in his lab. I have learned so much about DNA repair from both current and classic publications and the key concepts we studied in the journal club together. I would also like to thank Dr. David Green for accepting my offer for being in my thesis committee.

The Simmerling group also substantially contributed to the development of this work. Especially the strict and extensive comments and the many discussions and the interactions with Carlos Simmerling had a direct impact on the final form and quality of this thesis. Carlos was even available on his vacations from where he provided valuable comments on the work which led to this thesis. I would like to thank Dr. Viktor Hornak, Dr. Xiaolin Cheng, Dr. Guanglei Cui, Dr. Raphael Geney, Dr. Asim Okur, Dr. Bentley Strockbone, Dr. Daniel R. Roe, Salma B. Rafi, Lauren Wickstrom, Melinda Layten, Arthur Campbell, Christina Bergonzo, and Fangyu Ding, for our many discussions and providing me brotherly advises and tips that helped me a lot in staying at the right track. All work and no play make Ken a dull boy. I can not forget to mention the excellent time I have had playing Warcraft in the off-times with some of the above.

I had the pleasure to supervise and work with several high school students who did their summer work in our project and have been somehow beneficial for the presented work in this thesis. Karthik, Stephanie and Catherine have spent countless time and patience to build several new molecular models which made my several studies possible. The results of the various simulations have been used in chapter 5 of the thesis.

I am also grateful for the department of Chemistry at Stony Brook University for providing me an excellent work environment during the past years. I would like to thank the chemistry staff members, especially student affairs coordinators, Diane Godden and Katherine M. Hughes.

I am grateful for Dodi Heryadi who helped me to understand the NCSA supercomputing center. He was always patient and answered all my emails promptly with very helpful information.

I feel a deep sense of gratitude for my father and mother who formed part of my vision and taught me the good things that really matter in life. The happy memory of my days in China with my family still provides a persistent inspiration for my journey in this life. I am grateful for my Grandpa and Grandma, for rendering me the sense and the value of brotherhood and family. I am glad to be part of this family.

A special thank goes to Salma, whom I have known for more than 5 years now and who showed to be a kind, mostly helpful and trustful friend. I am very grateful for to her, for her love and patience during the PhD period and I hope she will remain my best best best best friend for the rest of my life. She has been the most annoying and persistent

person who has been with me in my good and bad times. I also want to thank many other friends I made over the passed years, Eric, Ron, Anna, Emma, Xiaojing, Liwen, Luming, Debbie, Andi, Sunyi, Chenjin, Liying, Chenlin, Huangqing, Joe, Aiwu, Sunliang, Xiuzhe, Zhangjun, Wangjie, and many many others. I wish you all the best in the future.

Chapter 1 Introduction

1.1 DNA damage and repair

1.1.1 DNA and DNA duplex

Deoxyribonucleic acid (DNA) is a biopolymer which contains genetic information of living organisms. All known cellular life forms and some viruses contain DNA. It contains information needed for constructing the rest of the living organism, such as proteins. The part of the DNA with this type of information is called a gene. It has been estimated that human genome includes 20,000 – 25,000 genes(*1*). Human genome project, a 13-year international project completed in 2003, has identified 92% of all genes in human DNA(*1*).

Although DNA contains a large amount of information, its basic structure is rather simple. Chemically, a DNA strand is a long-chain polymer molecule. Each monomer is a nucleotide. A nucleotide has three components: sugar, phosphate, and base. The sugar is a pentose deoxyribose. The base is a heterocyclic base. Depending on the structure of the ring, the base is called pyrimidine, or purine. Pyrimidine base is a heterocyclic aromatic ring, containing four carbon atoms and two nitrogen atoms at 1 and 3 positions. Purine base is a pyrimidine ring joined with an imidazole ring. DNA consists of four bases:: adenine, guanine, cytosine, and thymine that share the same sugars and phosphates.

Adenine and guanine are purines; cytosine and thymine are pyrimidines. Figure 1-1 shows the structure of the nucleotide and four types of bases.

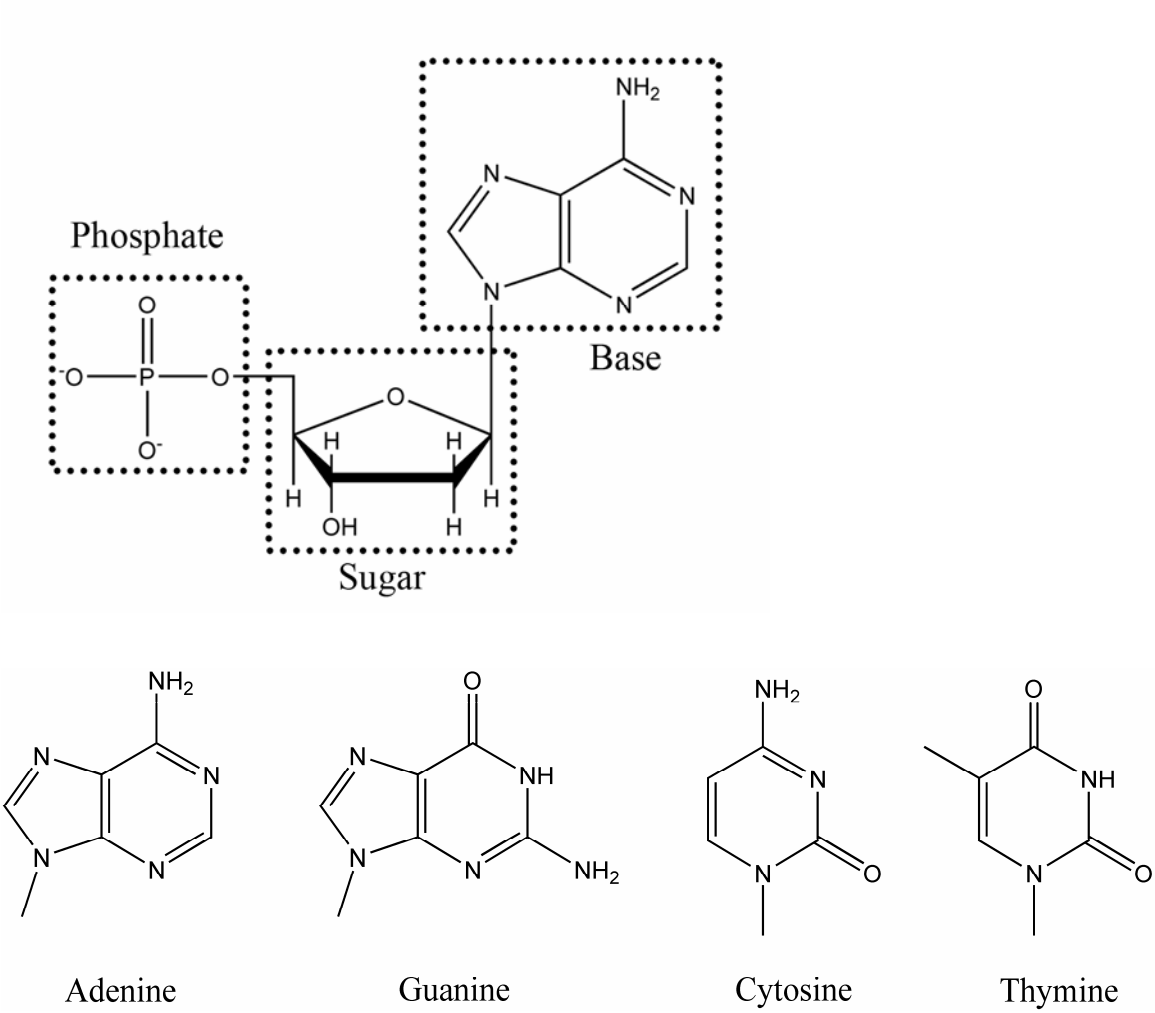

Figure 1-1: The structure of nucleotide, and the four types of bases. In the structure of nucleotide, adenine is used as an example of the base. The base can be any of the four types.

In vivo DNA is present as two complementary strands huddling, most commonly

assuming right hand helix conformations. The resulting structure is known as DNA duplex or double-stranded DNA. The discovery of DNA 3-D double helix structure by Crick and Watson has been one of the most significant breakthroughs in the history of life science(*2*). In this model the phosphates of the DNA are outside of the duplex and the bases are inside. The purine bases pair with pyrimidine bases by hydrogen bonds. Figure 1-3 is a demonstration of the DNA duplex structure in B-form, which is the most common form for DNA duplex *in vivo*. Adenine pairs with thymine and guanine pairs with cytosine. They form two or three hydrogen bonds between bases, which are called Watson-Crick base pairs (shown in Figure 1-2). This feature explains that the previous experimental observation that the molar ratios of the amount of guanine to thymine, and that of adenine to cytosine are close to the unity(*3*). This also implies the mechanism of the storage and transferring of the genetic information by DNA.

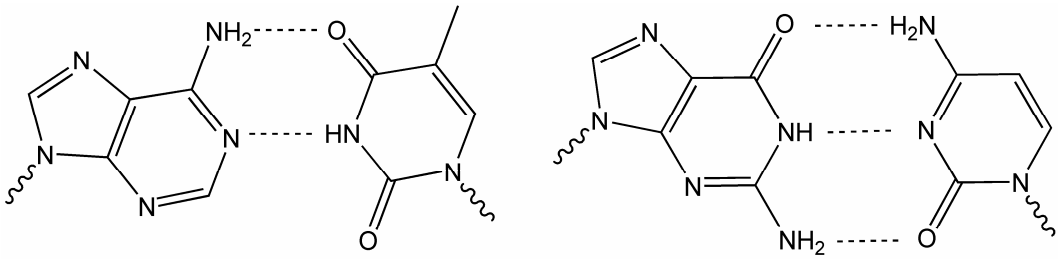

Figure 1-2. The two types of Watson-Crick base pairs in normal DNA.

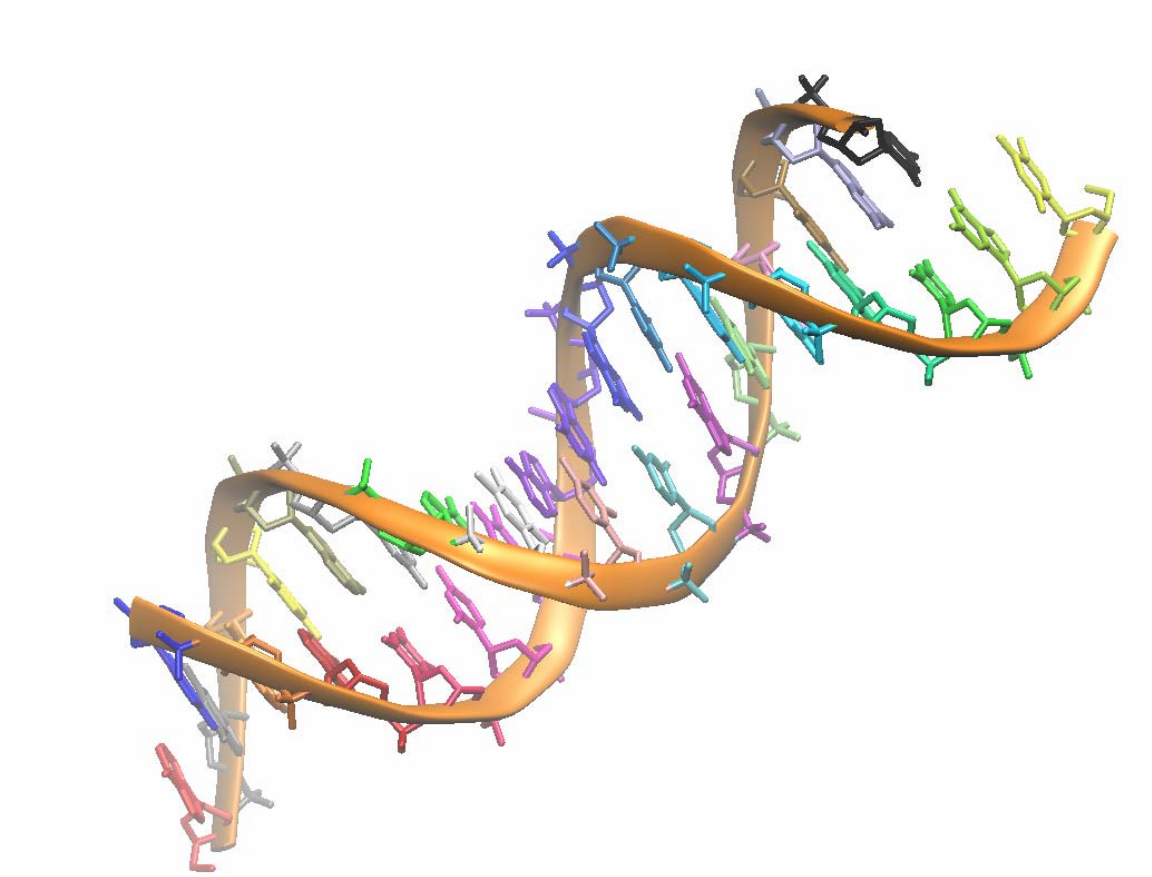

Figure 1-3. An example of the 3-dimensional DNA duplex structure in the B-form DNA, which is most common form *in vivo*. The backbone of the DNA is shown in orange ribbons. The heavy atoms of the nucleotides are shown in sticks. They are color coded according to the positions. We can see from this figure that all base groups are inside of the duplex, and pair with the base group from the complementary strand. The phosphate groups are on the surface of the duplex.

1.1.2 DNA damage and repair

Due to the high stability of DNA structure, it was two decades after Watson and Crick proposed the DNA duplex structure that the importance of DNA damage and repair was brought into serious considerations(*4*). This was much benefited from the realization of the life cycle of DNA as a dynamic process which includes the three "R's": Replication, Recombination, and Repair. It became apparent that DNA damage is common biological events to all living organisms. Damage can result from environmental factors such as UV, or endogenously, such as reactive oxygen species. It has been estimated that the frequency of DNA damage is about 1 million per cell per day(*5*). Fortunately, as Crick suggested that "DNA is so precious that probably many distinct repair mechanisms would exist"(*4*), all living organisms have evolved sophisticated DNA repair mechanisms and a large amount of proteins work in concert to protect the fidelity of DNA. In healthy cells, the rate of DNA repair is equal to the rate of DNA damage. In unhealthy cells, there is a DNA damage accumulation due to the inefficiency of DNA repair. One important focus in DNA repair field is to study the malfunctioning of the DNA repair enzymes which results in the accumulation of damaged DNA in cells.

There are three major DNA repair pathways: base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR)(*6*). BER recognizes the single nucleotide damage and excises the damaged base from the genome. NER is a more complex process. It recognizes bulky, helix-distorting lesions and excises DNA fragments containing about 30 nucleotides. MMR corrects errors of DNA replication and recombination that result in mispaired nucleotides. Single misplaced nucleotides are excised as a result of MMR

repair. After either type of the repair, the gap resulting from the excision is filled with new nucleotides in a reaction catalyzed by DNA polymerases. In this study we have focused on the guanine oxidative lesions, which are repaired through BER pathway by DNA glycosylase.

Before moving to the next section, there is another point to keep in mind. As we know that low mutation rates are necessary for stable life form(*7*), certain level of mutation is the fundamental reason for the genetic diversification and evolution. Indeed, "life is a delicate balance between genomic stability and instability – and of mutation and repair"(*8*).

1.1.3 8-oxo-guanine and GO system

Cellular DNA is constantly facing oxidative stress from both endogenous and exogenous sources. 8-oxo guanine is one of the most common mutagenic forms of DNA oxidative damage(*9, 10*). The structures of normal guanine and 8OG are shown in Figure 1-4. 8OG differs from guanine at the N7 and O8 positions in that N7 is protonated and the C8 hydrogen is replaced by oxygen.

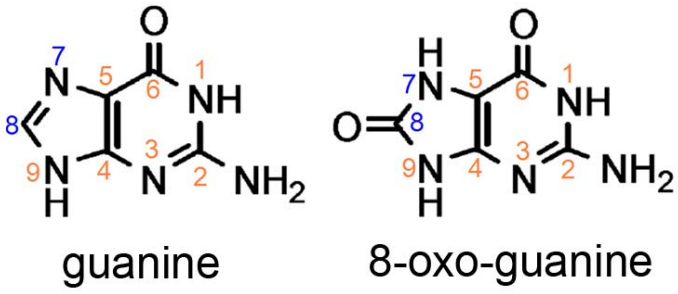

Figure 1-4: Structures of guanine and 8-oxoguanine.

8OG is especially deleterious because it can form two types of base pairs with both cytosine and adenine. Figure 1-5 shows the two base pair patterns 8OG can form. As does normal guanine, 8OG can formWatson-Crick base pairs with cytosine in its *anti* conformation (Figure 1-5A). 8OG can also form Hoogsteen base pair pattern with adenine in its *syn* conformation (Figure 1-5B)(*11*).

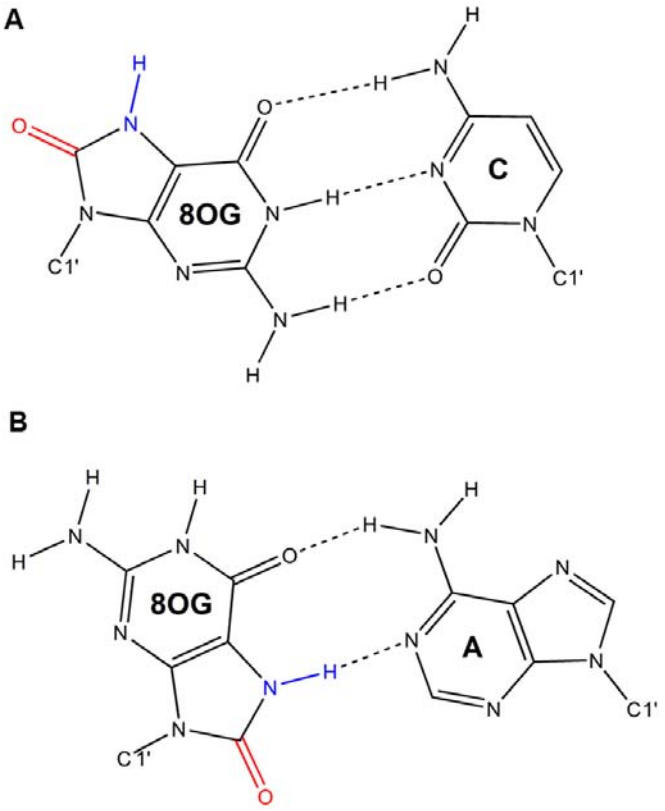

Figure 1-5. The two base pairing patterns of 8OG. Panel A shows the Watson-Crick base pair pattern which 8OG forms with cytosine. Panel B shows the Hoogsteen base pair pattern which 8OG forms with adenine.

Failure to repair the damaged base can cause G:C to T:A transversion (Figure 1-5), and studies have shown that this transversion mutation is highly related with aging and diseases such as cancer(*12-14*). In *Escherichia coli* and other bacteria, this lesion is repaired through a mechanism called the GO system, which includes the enzymes formamidopyrimidine glycosylase (Fpg, also called MutM), MutY and MutT(*15, 16*).

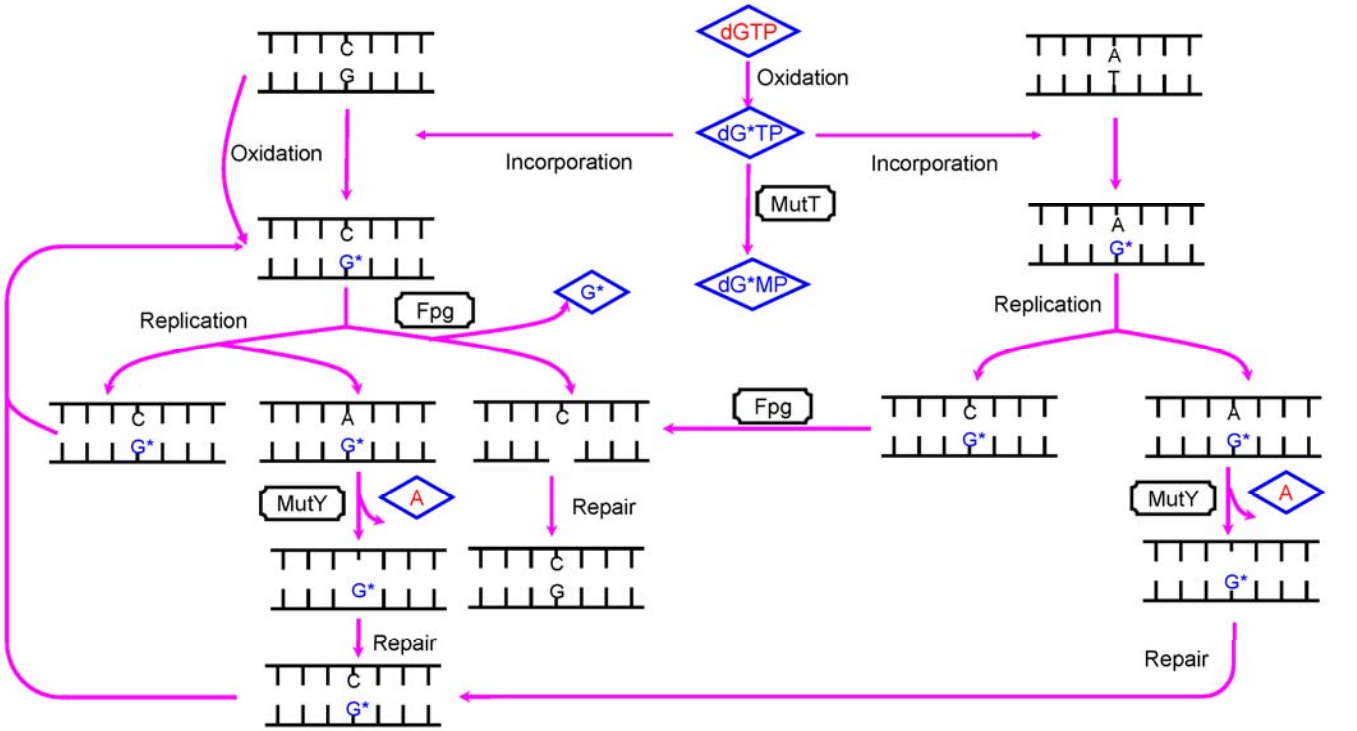

Figure 1-6. The scheme of GO system in *E. coli*. The functions of three enzymes, Fpg (also known as MutM), MutY, and MutT are working in concert to repair 8OG lesion. The G* represents 8OG.

Figure 1-6 shows how these three enzymes work together to repair 8OG oxidative damage in GO system in *E. coli*. The 8OG in the DNA duplex comes from two resources. The normal guanine in DNA duplex can be oxidized to 8OG. Also, dGTP can also be oxidized to 8OGTP and then incorporated into DNA duplex. To prevent the second condition, MutT decomposes the oxidized dGTP to dMTP, which cannot be used in DNA polymerization. For 8OG paired with cytosine, the enzyme Fpg recognizes the lesion and excises the oxidized base. If Fpg fails to capture the lesion before the replication cycle, 8OG can forms base pairs with both cytosine and adenine as shown in Figure 1-5. When 8OG pairs with cytosine, it can still be repaired by Fpg. When it pairs with adenine, the situation is more complicated. If 8OG is excised, the G to T transversion will become permanent. In this case, another DNA glycosylase, MutY comes into play. It recognizes the 8OG:A base pair, excises the adenine base, leaving the 8OG intact. Afterwards, a cytosine is filled opposite 8OG. The new 8OG:C can now be repaired by Fpg. Thus, MutY here acts as the "failsafe" mechanism for Fpg.

1.1.4 Fpg and its recognition mechanism on 8OG

Fpg as a DNA repair enzyme was first found in extracts of *E. coli* by Chetsanga and Lindahl in 1979. They reported that Fpg repairs alkylation lesions and damaged purine with open imidazole ring(*17*). The second function led to the assignment of the name "formamidopyrimidine glycosylase". Ten years later, the *mutM* gene was found in *E. coli* by Cabrera *et. al.* from Miller group at UCLA. This mutator strain has a high rate of G:C

to T:A transversion(*18*). Only after 3 years that Tchou and Grollman, discovered that Fpg is the product of the MutM gene and demonstrated to excise 8OG as well as damaged purines with open imidazole rings(*19*).

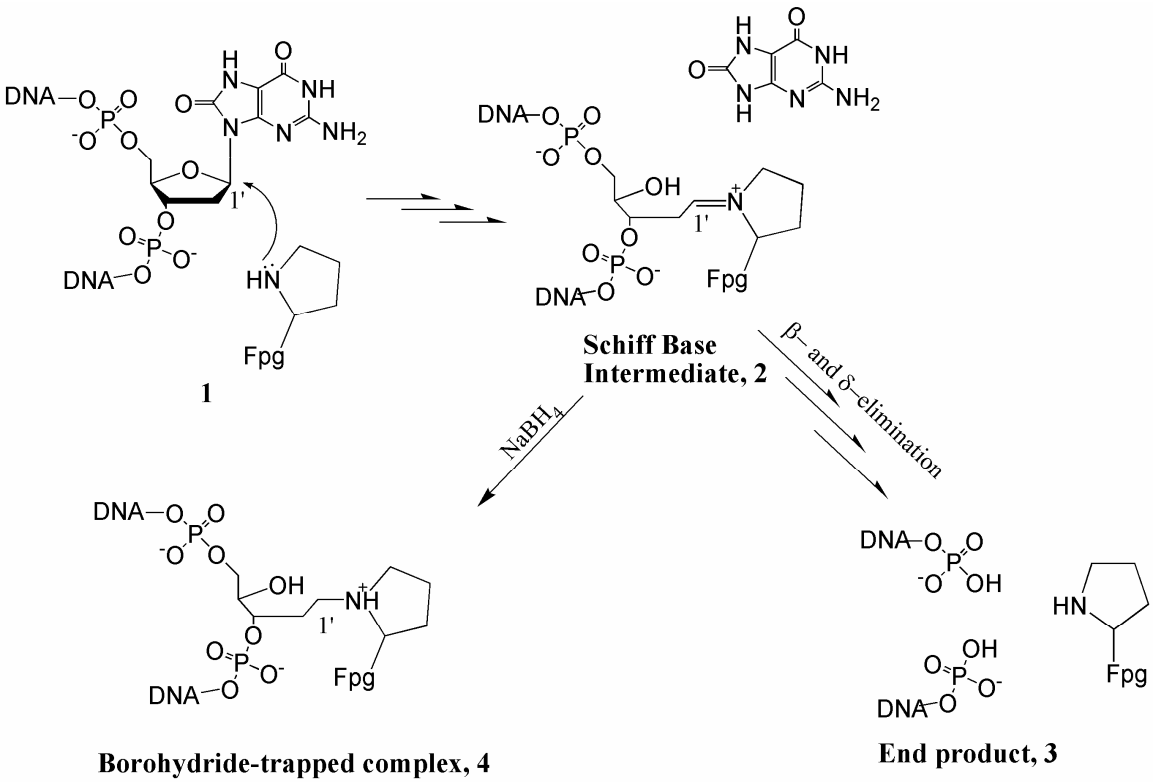

Figure 1-7. The proposed mechanistic scheme for the excision reaction catalyzed by Fpg (*20*). The structures show four stages of the excision. Stage 1 shows the scene after 8OG binds to the active site of Fpg. The catalytic residue Pro1 of Fpg attacks the C1' position of 8OG. This reaction and the following reactions result in the Schiff base intermediate, which is shown in stage 2. The base is completely excised from the sugar. Stage 3 shows the end product after a complete removal of the damaged nucleotide by β- and δ-eliminations.

Stage 4 is the covalent complex formed by trapping Schiff base intermediate using $NaBH_4$.

The function of Fpg in GO system is to recognize 8OG when it is paired with cytosine and to excise the lesioned nucleoside from DNA strand. Excision is a complex multi-step procedure, initialized by nucleophilic attack of the lesion C1' atom by the enzyme's N-terminal proline (1 in Figure 1-7), and formation of a covalent bond (2 in Figure 1-7) (*20, 21*). The resulting ring-opened Schiff base(*22*) subsequently undergoes β- and δ-eliminations (*23*), resulting in the complete removal of the damaged nucleotide(3 in Figure 1-7). The Schiff base intermediate can be reductively trapped by treatment with $NaBH_4$ and form a stable covalent complex (4 in Figure 1-7)(*20, 21*). The existence of the Schiff base was firstly proposed by Tchou *et al.* in 1994(*24*), and later confirmed using the reduction reaction using $NaBH_4$ (*20, 21*). This reaction was also used by Zharkov et al in generating covalently bonded protein/DNA complex for crystallography studies(*25-29*).

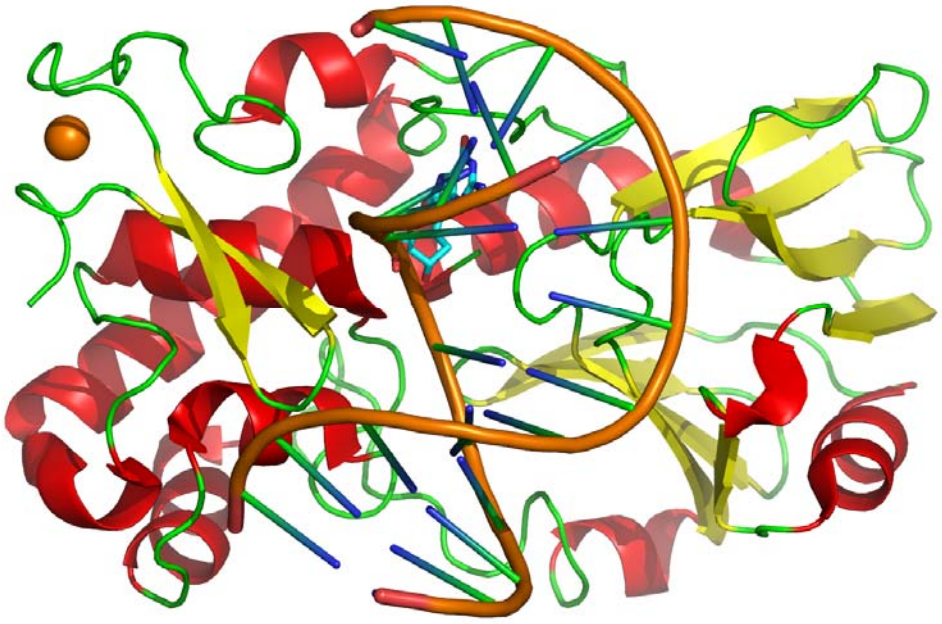

Figure 1-8. The cartoon representation of the x-ray structure of Fpg bound with 8OG containing DNA duplex (pdb id: 1R2Y). The secondary structures of Fpg are shown in red (α-helix), yellow (β-strand) and green (turn). The left half of the current view is C-terminal domain, which is rich in α-helices. The right half is N-terminal domain, which is rich in β-strands. The zinc finger motif is in the up-left corner in the current view.

Thanks to the work of x-ray crystallographers, a series of x-ray structures of Fpg under different conditions have been solved. A partial list is shown in Table 1-1. Fpg has about 270 residues, the exact number varying according to the species. Figure 1-8 shows an x-ray structure of *Bacillus stearothermophilus* Fpg/DNA complex. Fpg contains two distinct domains, linked by a flexible hinge region. A zinc finger motif, which is common

in DNA binding proteins, is in the C terminal domain. The binding loop, βF-α10 loop, is present above the active site. In the x-ray structures with 8OG present, the O6 atom of 8OG forms 4 hydrogen bonds with the N atoms from the 4 residues at the tip of the binding loop. It is interesting to see from Table 1-1 that the binding loop is present in x-ray structure of free Fpg, and Fpg/DNA complex with 8OG. However, its electron density is missing the x-ray structures of Fpg/DNA complex without 8OG. One possible explanation is that the binding loop is ordered in the free, relaxed structure of Fpg. When the Fpg binds to DNA, the two domains approach one other. The stress destabilizes the loop, which may also help to release the excised base. In the structures with 8OG in the active site, the hydrogen bond network between 8OG and the loop help to stabilize the base, also keep the flipped-out 8OG in the active site for further chemical reaction to occur. Further study is needed to exam this theory.

	8OG E3Q 1R2Y	8OG xlink 2F5Q, 2F5S	A:T xlink 2F5P	G:C xlink 2F5O	Free 1EE8	Abasic WT 1L1T, 1L2C, 1L2D
Lesion	Active site, O6 top multiple backbone H in βFα10 loop	Active site, same as no xlink	Intrahelical	Intrahelical	---	---
Binding βF-α10 loop	Ordered, same as unbound Fpg	same as no xlink	Missing	Missing	Ordered but not quite same as 8OG complex	Missing.
T224	Hbond to N7	Hbond to N7	Loop missing	Loop missing	Serine, loop moved	Loop missing
F114 (wedge)	present	same as no xlink	Present, retracted some	Same as 8OG, tilted slightly diff due to presence of G. 8OG one moves to hole a little.	Same location, rotated 180deg chi2.	Nearly identical to 8OG
M77 (wedge)	Plugs from minor groove side	same as no xlink	Moved completely (7-8A) outside minor groove to entry of active site, near p0.	In minor groove in plane of G:C near where it needs to plug. Close to unbound but retracted some.	In minor groove near where it needs to plug	Nearly identical to 8OG
R112	Plugs from minor groove side, HB to C	same as no xlink	In minor groove parallel to backbone- interacts with E78 in active site!	In minor groove parallel to backbone- interacts with E78 in active site!	Nearly same as 8OG bound	Plugging Nearly identical to 8OG

Table 1-1. The available x-ray structures containing Fpg and the comparison of the key features.

Before the lesion can be excised, Fpg has to recognize the damaged base and form the Michaelis complex (1 in Figure 1-7). In the present context, lesion recognition refers to the process by which a specific lesion, 8OG, is selected for excision by Fpg in a vast sea of unmodified DNA. The problem may be considered in two parts; how the enzyme "finds" the lesion embedded in DNA and how the modified base, once encountered, is accommodated as a Michaelis complex in the enzyme's active site. The current model for damage recognition involves a) non-specific binding to duplex DNA b) scanning the groove(s) by facilitated diffusion c) damage recognition mediated by hydrogen bonds or thermodynamic instability d) formation of a transient enzyme-DNA complex e) eversion of damaged nucleotide from helix and f) binding in the active site pocket (*30*). However, this model is currently only supported by the stop-flow fluorescence experiments, which does not provide structural information, This study is dedicated to simulate the lesion recognition process by breaking the whole procedure into individual steps according to this model. So we will be able to observe the all-atomic-resolution dynamics of the lesion search and eversion steps and compare the behaviors of 8OG and normal guanine. The differences observed will provide novel insights into how enzyme Fpg efficiently and accurately recognizes 8OG among a great excess of undamaged DNA.

1.2 Molecular dynamic simulation

Molecular dynamic (MD) simulations are widely used in structural biology and structure based drug design(*31-33*). Comparing with conventional experimental methods, MD simulations have several advantages. First, it exhibits biological events occurring on very short time scale. Unlike many experimental structural biology methods, which only generate time-averaged results, the timescale of MD simulation can be infinitely small, usually between femtoseconds to picoseconds. Second, MD simulation provides an approach to directly observe the biological events at atomic resolution. The time-sequence trajectories generated by MD simulation exhibit the atomic motions with the aid of graphic software such as MOIL-view(*34*), pymol(*35*) and VMD(*36*). This is extremely important for fast events that cannot be revealed directly by any other methods. Third, MD simulation can be used to explain biological events not only qualitatively, but also quantitatively. A relatively new concept, “iso-structure does not necessarily imply iso-energy”(*37*), is now generally accepted in the structural biology community. The systems with same or similar structures can be studied using the computational approaches to evaluate the energy differences. Several free energy calculation methods based on MD simulations, for example MM-PBSA(*38, 39*), thermodynamic integration (TI) (*40*), and free energy perturbation (FEP) (*41*) methods, can directly calculate the free energy of a system or free energy difference between systems. The goal of my research is to use MD simulations to simulate the biological events, explain the relevant experimental results from microscopic view, and understand the mechanisms of enzymes from both structure and energy points of view.

1.2.1 Basics, history and perspectives

Although MD simulations have been widely used as an essential tool for biological and other systems, its fundamental principle is rather simple. The particles in the systems are defined as spheres. Their movements obey Newton's three laws of motion. Before the collision, particles move with constant velocity. After the collision, the change of the velocity is directly proportional to the force acting on the particle and inversely proportional to the mass of the particle. During the collision, to every force applied there is an equal force applied in the opposite direction. In real applications, more smooth functional forms are usually used to replace this hard-ball model.

MD simulations were first introduced by Alder and Wainwright in 1957, and applied on hard sphere model to study phase transitions(*42*). The first molecular dynamics simulation of a realistic system was done by Rahman and Stillinger in their simulation of liquid water in 1974(*43*). The first protein simulations appeared in 1977 with the simulation of the bovine pancreatic trypsin inhibitor (BPTI)(*44*). Although the simulation was short (9.7 ps) and in vacuum, the results showed that the biomolecules are dynamic systems, unlike the static structure shown by x-ray crystallography.

Since first applied to biomolecular systems 30 years ago, MD simulations have been widely used in studies of biomolecules. One common pattern in applying MD in biological systems is using structural experimental methods, such as NMR or x-ray crystallography to capture static snapshots, or kinetic experimental methods to obtain macroscopic data. Then MD simulations are used to fill the trajectories between

structural snapshots or reproduce the macroscopic data to explain the results using microscopic structures. It is always important to compare results of MD simulations with relevant experimenta; data. Numerous MD simulations have been applied on complex biological problems such as protein folding, ion channel, motor protein function, and enzyme catalysis etc (*31, 32*).

Despite countless efforts to improve the efficiency and accuracy of MD simulations, three challenges remain for conducting a successful MD simulation: force field, searching and sampling.

The accuracy of the force field is essential for the accuracy of the simulation. Force field is an empirical approach to calculate the interactions between atoms. The parameter set of a force field is usually generated by fitting on the quantum calculation results or experimental results. The performance of a force field can be biased towards the original conformation. A well known example is that most commonly used force field, amber94, which has a strong bias favoring α-helix conformation(*45*). Generalized Born (GB) model, an implicit way to simulate solvent effect, is usually considered as a part of force field function. It has been demonstrated that GB model has the tendency to over-stabilize salt bridges(*46*). Therefore, careful evaluation of the force field on the subject system before production simulation is essential.

In spite of the increasing computing power, searching for alternative stable conformations besides the initial structure or native structure remains a challenging problem. It has long been realized that many biomolecules could have more than one

stable conformation. Many conformational changes are also correlated with enzyme functions. However, these stable conformations are separated by energy barriers. Overcoming these energy barriers and searching for alternative conformation become a major challenge in studying protein function by computational methods. Simulated annealing(*47*), softcore potential(*48*), targeted MD and other methods have been used to conquer this problem.

Sampling is a step further than searching. We can use a simple 2-state-mode system as an example. Starting from one state, once simulation reaches the other state, it can be called a successful searching. Sampling is for generating the Boltzmann populations for the two states. The relative populations of the two states are determined by their free energies, and at equilibrium state, they are constant. A successful sampling needs multiple transitions between the two states, and "correct" potential energy for both states.

The constant increasing computing power is drawing a more and more promising perspective for biomolecular MD simulations. The increase of computing power consists of two components. One is the evolution of the computer hardware. According to the average over past few decades, the hardware computing power increases by a factor of 10 about every 5 years. The other one is from the new, more efficient computing algorithms. Table 1-2 shows the speedup of the Amber program, which is one of the mainstream simulation software packages. The current algorithm is more than twice as fast as it was 6 years ago. All these forces have led the simulations of larger systems and longer timescale feasible. Recently an all-atom explicit solvent simulation of the complete satellite tobacco mosaic virus has been carried out for more than 50 ns(*49*). The system

includes roughly one million atoms. A survey about the MD simulations in the past five decades and a prediction of the possible simulations feasible in the future were made by van Gunsteren *et al.* recently (*31*). According to the current trend of the increasing computing power, at 2080 we will be able to simulate biomolecules in real timescale (currently the speed of the simulation is ~ 10^{17} times slower than real timescale), and at 2172 we will be able, in principle, to run all-atom simulations on human body!

Code	Release date	speed, ps/day
Amber 4.1	June, 1995	103
Amber 5	November, 1997	104
Amber 6	December, 1999	121
Amber 7	March, 2002	135
Amber 8	March, 2004	179
Amber 9	March, 2006	249

Table 1-2. A standard Amber benchmark, but over about a decade of code changes. The benchmark is "jac", which is dihydrofolate reductase (159 residue protein) in TIP3P water (23,558 total atoms). The data are available from Amber webpage (http://amber.scripps.edu).

Along with the rapid growth of computing power, great efforts have been made to improve the accuracy of the current force field and to design more sophisticated force fields. For example, in the MD simulations, atoms are considered as the elementary

particles and no electrons are included, in order to describe the enzymic reaction we will need hybrid quantum-classical (QM/MM) modeling. At the classical MD level, the polarizable force field, which includes terms to allow the polarization of the charge distribution according to the environment, is under development(*50*).

1.2.2 Force field

Force field is the mathematical description of physical interactions within a system. The force field used in classical molecular dynamics specifies how an atom interacts with the rest of the system. Simply, a force field is an equation with a set of parameters. It uses the nuclear positions of the atoms to calculate the potential energy of the system, usually with some additional terms to improve the accuracy of the output.

There are four components in the Amber force field: bond energy, angle energy, torsion energy, and non-bonded interaction energy. Two different force fields can have different parameters or functional forms.

$$V = \sum_{bonds} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{angles} \frac{k'_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{torsions} \frac{V_n}{2}(1+\cos(nw-r))$$

$$+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} (4\varepsilon_{i,j}[(\frac{\sigma_{i,j}}{r_{i,j}})^{12} - (\frac{\sigma_{i,j}}{r_{i,j}})^{6}] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}) \qquad 1.1$$

Equation 1.1 is the general form of the Amber force field (*51*). The first term is the bond term which models the interactions between pairs of bonded atoms. The second term is the summation of all the valence bond angle's bending. These two terms are

modeled by the harmonic potential, k_i and k'_i are the force constants, and l_i-$l_{i,0}$, Θ_i-$\Theta_{i,0}$ are the deviations from their respective equilibrium values. The third term is the torsional potential that models how the energy changes as a bond rotates. The fourth term models the non-bonded interactions which are usually modeled by using the Coulomb potential term for electrostatic interactions and a Lennard-Jones potential for van der Waals interactions. Usually the largest difference between two different force fields comes from the philosophies for optimization of the non-bonded parameters. An illustration of these four terms is shown in Figure 1-9.

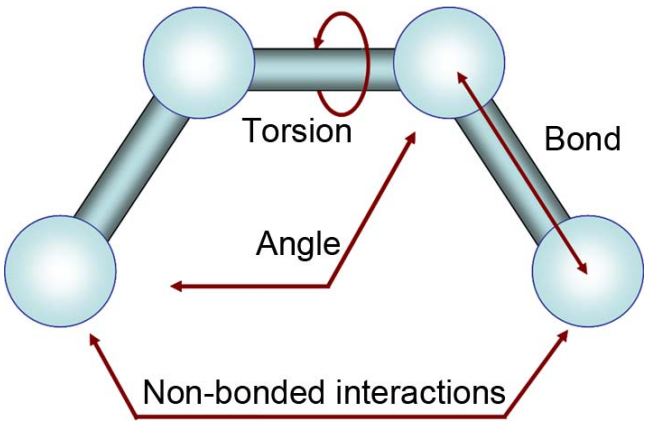

Figure 1-9. Four types of interactions in amber force field.

One of the most important qualities of a "good" force field is that its functional form and parameters must be transferable. This means that the same set of parameters can be used to model a series of related molecules. However, careful evaluation of the accuracy

of the force field on the subject before further analyzing simulation results is always critical.

1.2.3 Solvation effects

In most cases we are interested in the properties of molecules in solution, usually in aqueous solution. Therefore, it is essential to calculate not only the interactions between the atoms of biomolecules, but also the solvent effects on these interactions. There are two different approaches to include the solvent effects in molecular dynamics. One is the explicit solvent model, the other one is the continuum solvent model, also known as the implicit solvent model.

In explicit solvent model the solvent molecules are explicitly included in the calculation. One of the most commonly used models is the TIP3P water model [16]. In explicit solvent models we need to calculate the interaction between the particles and every solvent molecule individually, therefore it is a very "computationally expensive" job. Algorithms such as Particle Mesh Ewald method(*52*), particle-particle/particle-mesh (P3M) method(*53*) etc. are developed to accelerate the simulations.

Even with the advanced algorithms, one obvious drawback of explicit solvent model is the large system size due to numerous solvent molecules. An alternative approach is using the continuum solvent models, such as the Generalized Born (GB) model. This model represents the solvent implicitly as a continuum with the dielectric properties of water, and also includes the charge screening effects of salt. This lowers the calculation expense in two aspects: first, it removes the calculation of the interactions and motions

involving solvent molecules; second, the absence of solvent friction accelerates the dynamics of solute. Equation 1.2 shows how to calculate the solvation energy using GB model by considering a process in which the molecule is transferred from a uniform medium of dielectric value D_p into a solvent with dielectric constant D_w. There are two terms in it. The first term is to calculate the energy of creating the charge distribution in a uniform dielectric D_p, and the second term to calculate the solvation energy. f^{gb} is a certain smooth function which is assumed to depend only upon atomic radii ρ and inter-atomic distances r_{ij}.

$$E_{i,j} = \frac{1}{2}\sum_{i \neq j} \frac{q_i q_j}{D_p r_{ij}} - \frac{1}{2}\left(\frac{1}{D_p} - \frac{1}{D_w}\right)\sum_{i,j} \frac{q_i q_j}{f^{gb}(r_{ij})} \qquad 1.2$$

1.3 Advanced MD simulation methods

As mentioned earlier, due to the rugged potential surface of biomolecules and the limited computing power, advanced sampling methods based on MD simulations are commonly needed to enhance sampling. In this section several sampling methods used in my research are introduced.

1.3.1 Targeted molecular dynamics

One way of generating widespread conformations is Targeted Molecular Dynamics (TMD), in which a simulation is initiated in the structure at one endpoint of the path and a biasing potential slowly forces the system to adopt the other endpoint (*54*,

55). In O\our applications, an additional term was added to the energy function, named as the targeting force. The targeting force is calculated based on the equation 1.3 with a reference structure and a target RMSD value which are given as additional input for the simulation.

$$E_{tf} = K \times N \times (RMSD - RMSD_0)^2 \qquad 1.3$$

Equation 1.3 shows the calculation of targeting force. Here K is the force constant and N is the number of atoms. $RMSD$ is the structural difference between current snapshot and the reference structure. $RMSD_0$ is the target RMSD value. When the calculated best-fit RMSD value differed from the targeted value, the atomic derivatives of this term forced the system toward or away from the target (depending on the sign of K).

TMD has been widely applied to the study of large conformational changes in biomolecules (*56-61*).

1.3.2 Replica exchange molecular dynamics

The potential energy surfaces of the complex biological systems are rugged with various local minima. Simulations under normal temperature can be trapped in these local minima and preclude success even when a sufficiently accurate Hamiltonian of the system is used in the simulations. One simple solution is using high temperature simulations, using the high kinetic energy to escape these local minima. However, the

properties generated from the high temperature simulations are usually not biologically relevant. Replica exchange molecular dynamics (REMD)(*62*, *63*), also known as parallel tempering(*64*) simulations combine both high temperature simulations and low temperature simulations into one calculation to increase the efficiency of sampling under physiological conditions.

In standard Parallel Tempering or Replica Exchange Molecular Dynamics, the simulated system consists of *M* non-interacting copies (replicas) at *M* different temperatures. The positions, momenta and temperature for each replica are denoted by ($q^{[i]}$, $p^{[i]}$, T_m), i = 1,…,M ; m = 1,…, M. The equilibrium probability for this generalized ensemble is

$$W(p^{[i]},q^{[i]},T_m) = \exp\{-\sum_{i=1}^{M}\frac{1}{k_B T_m}H(p^{[i]},q^{[i]})\} \qquad 1.4$$

where the Hamiltonian $H(p^{[i]},q^{[i]})$ is the sum of kinetic energy $K(p^{[i]})$ and potential energy $E(q^{[i]})$. For convenience we denote $\{p^{[i]},q^{[i]}\}$ at temperature T_m by $x_m^{[i]}$ and further define X = $\{x_1^{[i(1)]},...,x_M^{[i(M)]}\}$ as one state of the generalized ensemble. We now consider exchanging a pair of replicas. Suppose we exchange replicas *i* and *j*, which are at temperatures T_m and T_n respectively,

$$X = \{...; x_m^{[i]};...; x_n^{[j]};...\} \rightarrow X' = \{...; x_m^{[j]};...; x_n^{[i]};...\} \qquad 1.5$$

In order to maintain detailed balance of the generalized system, microscopic reversibility has to be satisfied, thus giving

$$W(X)\,\rho(X \rightarrow X') = W(X')\,\rho(X' \rightarrow X) \qquad 1.6$$

where ρ(X→X') is the exchange probability between two states X and X'. With the canonical ensemble, the potential energy *E* rather than total Hamiltonian *H* will be used simply because the momentum can be integrated out. Inserting equation 1.4 into equation 1.6, we obtain the following equation for the Metropolis criterion for the exchange probability:

$$\rho = \min\left(1, \exp\{(\frac{1}{k_B T_m} - \frac{1}{k_B T_n})(E(q^{[i]}) - E(q^{[j]}))\}\right) \qquad 1.7$$

In practice, several replicas at different temperatures are simulated simultaneously and independently for a chosen number of MD steps. Exchange between a pair of replicas is then attempted with a probability of success calculated from equation 1.7. If the exchange is accepted, the bath temperatures of these replicas will be swapped, and the velocities will be scaled accordingly. Otherwise, if the exchange is rejected, each replica will continue on its current trajectory with the same bath.

REMD method has successfully applied to describe the free energy landscape of peptides(*65-68*) and proteins(*69, 70*), the amyloid formations(*71, 72*), and ligand binding(*73*).

1.3.3 Umbrella sampling method

Umbrella sampling is used to calculate the free energy profile along the reaction coordinate (*74-77*). In umbrella sampling, a series of simulations are performed using the

true potential energy function plus a biasing potential. The biasing potential is designed to introduce the strength of the bias along a "reaction coordinate" to sample regions on the "reaction coordinate" that may not otherwise be explored extensively. By doing so, each simulation will sample the conformation space, which is like an umbrella formed by the biasing potential. Post-processing using weighted histogram analysis method (WHAM) can be used to eliminate the systematic bias in a formally exact manner.

$$A(x) = -k_B T \ln P'(x) - U'(x) \qquad 1.8$$

Equation 1.8 shows the general form of the equation for calculating unbiased free energy. $P'(x)$ is the population sampled with biased potential. $U'(x)$ is the biased potential. k_B is the Boltzmann constant. T is the temperature.

For umbrella sampling simulations, three key parameters have been chosen carefully. One is the reaction coordinate. The final free energy profile is the projection of the potential of the system on the reaction coordinate. Therefore, the reaction coordinate should be able to represent the physical properties that the researchers are interested in. It also should be feasible to apply biased potential along the reaction coordinate. The other two values are the size of each window and the force constant for the biased potential. Umbrella sampling simulation consists of a series of restrained simulations. Each of the restrained simulation is called a window. They have the same simulation conditions, except the reference value, which is the location of this window on the reaction coordinate. The difference between two reference values of neighboring windows is the window size. Too small a windows size means more windows needed. Too large a

window size could cause the failure of the post-processing or induce error in the resulting free energy profile because there may not be sufficient overlap between neighboring windows. To construct the free energy profile, the WHAM analysis needs the adjacent windows have the same free energy value for the overlap region. Less than sufficient overlapping will introduce the statistical errors in each individual estimate. The force constant of the restraint simulations defines the location and range of the actual sampling. If the force constant is too strong, the range of the sampling would be small. More windows would be needed to generate enough overlap between windows. If the force constant is too weak, the restrain force will not be sufficient to create a structure ensemble in the desired region. For a successful umbrella simulation, all three values have to be chosen carefully, usually by trial-and-error fashion.

Umbrella sampling methods have been successfully applied to the calculation of strength of salt bridges(*46*), the relative binding free energy(*78*), and DNA base flipping(*79-83*).

1.3.4 MM-PB(GB)SA method

MM-PB(GB)SA method is another free energy calculation method(*39*). Unlike umbrella sampling methods, this is a post-processing method, which means it calculates the free energy based on existing structure snapshots. In MM-PB(GB)SA calculation, the free energy of the system is divided into three parts: molecular mechanic energy (MM), polar solvation free energy (PB or GB, depending on the method), and non-polar solvation free energy. The MM energy usually consists of electrostatic energy and van

der Waals energy of the system in vacuum. In some applications the bond energies, angle energies, and dihedral angle energies are also included in MM energy. The entropy of the system can be included by using normal mode analysis or other methods. But in most applications, MM-PB(GB)SA is used to calculate the relative energy between similar states, in which the entropy contribution to the total free energy will be canceled out. However, in the cases when the freedoms of the ligands are significantly different, the entropy term should not be ignored. One shortcoming with MMPB(GB)SA calculation is that it does not include the water molecules explicitly, which is problematic for the cases such as water bridge formation in the ligand binding(*84*).

1.4 Overview of my research

1.4.1 Structural Insights for alanine-rich model peptides: Comparing NMR helicity measures and replica exchange molecular dynamics in implicit solvent

The accuracy and precision are two key qualities of a successful simulation. The essential part of the accuracy of a simulation is the performance of the force field. The first part of my thesis is using a small model peptide to test the performance of the force field. The temperature dependence of helical propensities for the peptides Ac-GGG-$(KAAAA)_3$X-NH2 (X = A, K, and D-Arg) were studied using replica exchange molecular dynamics simulations and the simulation results are compared with data obtained from NMR chemical shifts of -$GG(KAAAA)_3X$-NH_2 and Ac-$(KAAAA)_3$XGY-

NH_2 sequences (X = A, K, and D-Arg). The chemical shift experiments were done in the laboratory of Dr. Niels Andersen at the University of Washington. A good agreement is found with both the absolute helical propensities as well as relative helical content along the sequence. Cluster analysis showed that the global minimum on the calculated free energy landscape corresponds to a nearly fully α-helical conformation. Energy component analysis shows that the single helix state has favorable intra-molecular electrostatic energy due to hydrogen bonds, and the globular states have favorable solvation energy. Furthermore, both experimental and simulation studies shown increasing helicity in the series X = Ala → Lys → D-Arg. The roles of these D-ending groups were analyzed in more details.

1.4.2 Computational analysis of the binding mode of 8-oxo-guanine to formamidopyrimidine-DNA glycosylase

8-oxo-guanine (8OG) is the most prevalent form of oxidative DNA damage. In bacteria, 8OG is excised by formamidopyrimidine glycosylase (Fpg) as the initial step in base excision repair. To efficiently excise this lesion, Fpg must discriminate between 8OG and an excess of guanine in duplex DNA. In this study, we explore the structural basis underlying this high degree of selectivity. Two structures have been reported in which Fpg is bound to DNA, differing with respect to the position of the lesion in the active site, one structure showing 8OG bound in the *syn* conformation, the other in *anti*. Remarkably, the results of our all-atom simulations are consistent with both structures. The *syn* conformation observed in the crystallographic structure of Fpg obtained from *B. stearothermophilus* is stabilized through interaction with E77, a non-conserved

residue. Replacement of E77 by Ser, creating the Fpg sequence found in *E. coli* and other bacteria, results in preferred binding of 8OG in the *anti* conformation. Our calculations provide novel insights into the roles of active site residues in binding and recognition of 8OG by Fpg.

1.4.3 Molecular Mechanics Parameters for the FapydG DNA lesion

FapydG is a common oxidative DNA lesion involving opening of the imidazole ring. It shares the same precursor as 8-oxodG and can be excised by the same enzymes as 8-oxo-guanine. However, the loss of the aromatic imidazole in Fapy-dG results in a reduction of the double bond character between C5 and N7, with an accompanying increase in conformational flexibility. Experimental characterization of Fapy-dG is hampered by high reactivity, and thus it is desirable to investigate structural details through computer simulation. We show that the existing Amber force field parameters for Fapy-dG do not reproduce experimental structural data. We employed quantum mechanics and molecular mechanics to calculate the energy profile for the rotation of the dihedral angles in the ex-imidazole moiety. Using these parameters, all-atom simulations in explicit water reproduce the crystallographic structure of *c*Fapy-dG in complex with the glycosylase Fpg. This result also confirms that the *c*Fapy-dG, which was used in the x-ray crystallography experiment, is a reasonable substitute for this type of studies.

1.4.4 DNA sliding and flipping in the Fpg/DNA complex

Sliding and base flipping are two essential motions in DNA lesion searching and recognition. However, the normal time scale of MD simulation (ns) cannot reproduce

these processes. In this study we used targeted MD and long MD simulations (1.6 ms in total) to simulate these two processes. R111 and F114 are two key residues for stabilizing the 8OG-flipped structure. In the base flipping simulations, R111 spontaneously entered the cavity created by the base flipping and formed hydrogen bonds with widowed cytosine. In DNA sliding simulations, F114 left the original position and entered the new position along with the DNA sliding. Both final structures are consistent with x-ray structures. In the long MD simulations, we observed the 8OG:C base pair breaking. Structural overlap showed that this semi-open DNA structure is very similar to the recent x-ray structure in which hOGG1 binds to the DNA. Multiple simulations starting form this intermediate state showed that this is a stable conformation. Two domain motions revealed by anisotropic network mode analysis also indicate the possible correlation between the structure and functions of Fpg.

1.4.5 The Role of Phe114 in Fpg in Searching and distinguishing Damaged DNA Base 8OG

The phenylalanine (give the residue number?) of Fpg present in the interrogating site and partially inserted in the DNA duplex has been observed in all available x-ray structures of Fpg/DNA complex systems. Besides the obvious stacking interactions with neighboring base groups, the role of this phenylalanine in lesion recognition has not been studied. In this study we calculated the free energy profiles of pulling the sidechain of this phenylalanine out of the duplex with 8OG:C base pair or G:C base pair in the interrogating site using umbrella sampling. The results showed that the energy barrier for pulling the phenylalanine with 8OG:C is higher than that with G:C base pair. Since

during the sliding process the wedge has to leave the pocket inside the duplex, this implies that sliding will be slower when 8OG:C is encountered, which will create a larger window for further conformation change, such as base pair breaking and base flipping, to occur. Detailed energy analysis showed that the difference of the energy barrier results from the repulsive interactions between O8 of 8OG and the neighboring phosphate atoms.

Chapter 2 Structural Insights for alanine-rich model peptides: Comparing NMR helicity measures and replica exchange molecular dynamics in implicit solvent

2.1 Introduction

Helix-coil theory is one of the most fundamental and intensively studied aspects of biomolecular structure(*85*). There are several reasons: first, the α-helix is a key secondary structure element in globular protein structure(*86*). Second, there are several experimental methods available which can detect the α-helical content. Third, the formation of α-helices are fast (sub-microsecond scale) relative to folding of proteins with multiple secondary structure units. This fast rate allows the all-atom simulations of α-helix formation and also makes helix formation a likely pre-equilibration event in folding of proteins with multiple secondary structure units.

Helix formation has long been known to show significant cooperativity effects, with coupling of hydrogen bonding and ϕ/ψ propensities between neighboring residues indicated by early theoretical(*87, 88*) and experimental(*89*) studies. The recent

investigations of amyloid fibrils have also found that the same sequences that adopt α-helices in globular proteins can form β-strands in amyloid aggregates(*90, 91*) The quest to understand these secondary structure transitions has brought helix studies into the spotlight.

Designed alanine-rich helices in which solubility and helical content is enhanced through inclusion of polar residues, are the most common short sequences used to study helix formation(*92*). Several simulation studies on helical propensity were reported recently. Zhang *et al.* studied the helix formation of two alanine-based peptides (Fs-21 and MABA-Fs) using standard MD simulations. They found that most populated structure for Fs-21 at room temperature is an α-helix bundle, while the low temperature ensemble is dominated by conformations with a single long helix(*93*). Nymeyer and Garica applied replica exchange molecular dynamics to the peptides A21 and Fs-21 using continuum and explicit solvent models. The result from the continuum model simulations predicted that the native structure of F-21 is a helix bundle. However, the explicit solvent simulations showed that the single helix structure is the most stable state. They proposed that this discrepancy was caused by the presence of incorrect hydrogen bonds stabilized by the continuum solvent model(*94*). A more recent study of helix formation was reported by Jas and Kuczera. In this study they used far UV CD spectra and replica exchange molecular dynamics to measure the helical propensities of the peptide Ac-WAAAH-(AAARA)3-A-NH_2. CD spectroscopy was used to calculate helical content and the melting profile. In their continuum solvent simulations, the global free-energy minimum was a single helix at low temperatures(*95*). These apparent discrepancies

between the different studies may arise from differences in peptide sequences, or may reflect the different simulation methods employed. In any case, a general model for helix formation has not arisen from these computational studies.

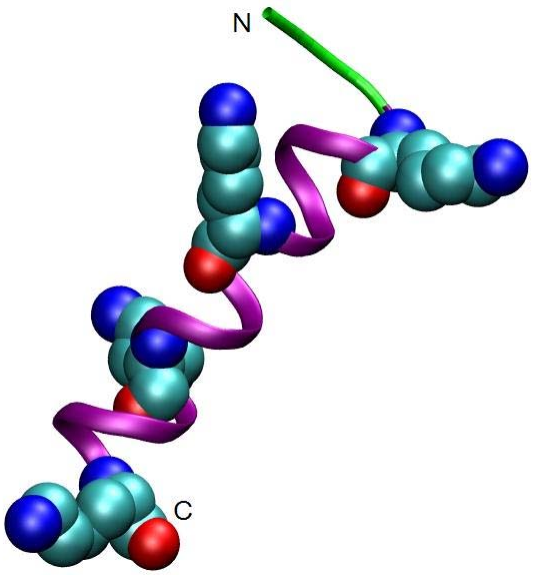

Figure 2-1. Helical conformation of Ac-GGG-(KAAAA)3-X-NH_2. The side chains of lysine are shown. The glycine residues are shown in green.

We report NMR and molecular dynamics simulation studies on a series of peptides, N-cap-(KAAAA)3-C-cap, closely related to the design previously reported by Marqusee *et al.*(*92*) (Ac-YGG-3X-K-NH_2 = Ac-YGG-(KAXAA)3-K-NH_2). The single helix conformation of the peptides and the position of lysine residues are shown in Figure 2-1. Contrary to the prior theoretical predictions that peptides of such a short length would not adopt measurable helix content in aqueous solution, Marqusee *et al*. demonstrated that peptides containing lysine and alanine can show as much as 80% helix in water (for X =

A). Other experimental studies have shown that the α-helix conformation is not favorable for short polyalanine peptides in water at room temperature(*96*), and several roles have been suggested for the stabilizing effects of polar side chains, particularly Lys and Arg(*67*). However, recent studies have provided propagation values for alanines (1.6) and a non-terminal lysine (0.82) that are in accord with those from the Baldwin group and not indicative of helix stabilization by the lysine side chain when it is not in the C-terminal turn of the helix(*97*). Positive charges at the C-terminus of a helix are expected to favor helix formation due to a favorable interaction with the helix macrodipole. D-Arg has been reported to be a particularly favorable C-cap(*98*). Huang *et al.* used static ultra far-UV circular dichroism and the time-resolved infrared spectroscopy coupled with laser-induced temperature-jump to detect the thermodynamic and kinetic properties of Ac-YGG-3A-D-Arg-NH_2(*99*). As a result, we also focus in this study on the contributions of the charged residues to the stabilization of the helix conformation. Explicit solvent simulation has suggested that charged side chains can favor helix formation by shielding the backbone hydrogen bonds from solvent(*67*). In this report, the results from all-atom simulations are compared to those obtained from NMR and CD experiments. We obtain strong agreement that provides validation of our simulation protocols and force field parameters. As well as an atomic-detail model for the structure ensembles that give rise to the experimental observables.

2.2 Methods

2.2.1 Replica exchange molecular dynamics simulations

Standard molecular dynamics simulations can become trapped in local minima during computationally affordable simulations. Therefore, replica exchange molecular dynamics (REMD)(*62, 63*) (also known as parallel tempering MD(*64*)) was used as an enhanced sampling method for this study. we used REMD as implemented in Amber version 8(*51*). REMD employs multiple non-interacting replicas of the system that are simulated independently and simultaneously at several different temperatures. At intervals of 1 ps, exchanges were attempted between conformations at neighboring temperatures based on a Metropolis-type criterion that considers the probability of sampling each conformation at the alternate temperature. The advantage of this method is that the simulations can escape from kinetic traps by "jumping" into alternate local minima being sampled more efficiently at higher temperatures. We employed 12 replicas at 253, 276, 300, 326, 355, 386, 420, 457, 497, 541, 588, and 640 K.

A modified force field based on AMBER parm94(*100*) was used to represent the peptide. Many Amber parameter sets have been reported to over stabilize α-helical structure(*45, 101*). In the parameters that we used, the torsional angle terms were adjusted to reduce this tendency. The generalized Born implicit solvent model(*102, 103*), with intrinsic Born radii taken from Tsui and Case(*104*) as implemented in Amber, was used to represent the effects of solvent. All non-bonded interactions were evaluated at every step. The SHAKE algorithm(*105*) was used to constrain the bonds involving hydrogen atoms. A 0.002 ps time step was used. Each replica was coupled to a constant-

temperature bath using a weak-coupling algorithm(*106*). The REMD simulations were carried out for 56 nanoseconds (a total of 672 ns simulation), and 56,000 snapshots were saved for each temperature.

2.2.2 Analysis of secondary structure content

To compare our results with previous studies, Lifson-Roig helix-coil theory(*107, 108*) has been used. In Lifson-Roig theory, there are three states for residues: the "l" state is the residue in random coil conformation, "w" is the residue in helical conformation, and its previous and next neighbor residues are in helical conformation as well, and "v" is a helical residue neighboring with one or two non-helical residues. Overall helical content corresponds to the number of w state residues, with a maximum helix length of N-2 for N residues. Following the work of Garcia and Sanbonmatsu(*67*) and Sorin and Pande(*109*), a residue is considered in a helical conformation if ϕ =-60(±30)° and ψ =-47(±30)°.

To complement this approach, we also used DSSP(*110*) (as implemented in the ptraj module of Amber) to analyze secondary structure content.

The melting curve is generated based on the temperature dependence of the average helical fraction (HF), calculated using the equation 2.1, where Nw is the number of w-state residues assigned by Lifson-Roig theory and N is the total number of residues.

$$HF = Nw/(N-2) \qquad 2.1$$

2.2.3 Free energy landscapes

We reconstructed free energy landscapes using principle component analysis (PCA), as implemented in ptraj in Amber 8, to define the reaction coordinates and 2-D histogram analysis to obtain relative free energy for each bin. A sample of input file for PCA analysis is shown in Appendix A – Principle component analysis. Due to their high flexibility, Gly residues were not included in this analysis. The alpha carbon coordinates of all other residues were used to generate the covariance matrix. The two largest eigenvalues from the PCA analysis were used to represent order parameters. The relative free energy of each bin was calculated based on the equation 2.2, where x is a histogram bin corresponding to a set of PCA eigenvector values, *G(x)* is the free energy relative to the most populated bin, *R* is the gas constant, *T* is the temperature, *P(x)* is the population in bin *x* and *P(0)* represents the population of the most populated bin.

$$G(x) = -RT\ln[(P(x)/P(0)] \qquad 2.2$$

2.2.4 Structure analysis

Cluster analysis was used to separate the ensembles into conformation families and representative structures. Moil-view(*34*) was used to perform the cluster analysis using backbone RMSD for non-Gly residues. Clusters were formed with a bottom-up approach using a similarity cutoff of 2.5 Å. The radius of gyration (R_{gyr}) was also used to

characterize structural properties of the overall ensemble and individual clusters. The carnal module of Amber was used for R_{gyr} calculation. All alpha carbon atoms of non-Gly residues were included in this analysis.

2.2.5 Energy decomposition analysis

To analyze the relative importance of various energy components in different structure families, energy decomposition analysis was carried out using post-processing in the sander module of Amber. Average total energies were separated into electrostatic, van der Waals and GB solvation terms.

2.2.6 Estimation of simulated data uncertainties

Lower bounds to data uncertainties were estimated as the absolute value of the difference between properties obtained using the entire ensemble sampled and that using only the second half of the data set.

2.3 Results and Discussion

2.3.1 Convergence test for REMD simulation

Although it has been shown that RMSD can greatly improve the efficiency of sampling, there is no quantitative analysis to show how long it actually takes for the results to converge. Before we started analyzing the simulation results, we first examined the convergence of the simulations.

The system peptides can have non-helical, single helix, or helix-turn-helix

conformations. We used the population of single helix conformation as the indicator. It is calculated by averaging the population of single helix over the range from 1 frame to the current frame.

Due to their similarity, the converging rates of the three peptides should be similar as well. Therefore, we only chose K19 for this test. The temperature is 275 K. The results are shown in Figure 2-2.

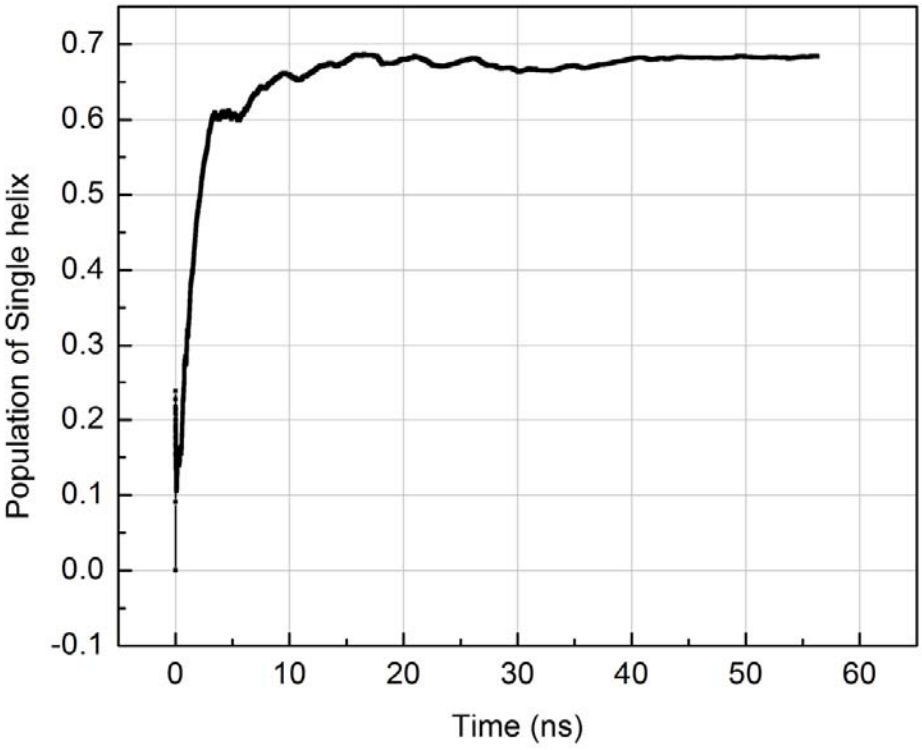

Figure 2-2. The rate of convergence. The x axis is the time. The y axis is the population of the single helix calculated by averaging over the region starting from 0 ns to the current frame.

As the initial structure is linear, the initial value at 0 ns is 0. Along with the

simulation processing, the population of single helix increased. At about 10 ns, the value reached 66% and stayed about that value for the rest of the simulations. The results reached the convergence at about 10 ns. Our simulation results are 5 times longer, which ensures that our simulation is precise.

2.3.2 Helical Propensities

Molecular dynamics simulations were performed for AcGGG-(KAAAA)3-X-NH_2 with X = Lys, D-Arg and Ala. Henceforth these are designated as peptides K19, DArg19 and A19. For all three peptides, the residues which are near the C-terminus have greatly reduced average helical content corresponding to the substantial helix fraying evident in the ^{13}C-NMR studies. The GGG units were non-helical ($f_H < 0.04$) with a rapid helicity increase beginning at K4 consistent with a significant N-capping effect by the GGG unit. The helical domain is viewed as K4 – A18/X19 (X19 can be either a helix C-cap or the frayed end of a helix). A greater extent of N-terminal fraying was evident for A19. For the other systems, C-terminal fraying was much greater than the fraying of the capped N-terminus. The simulations indicated a greater overall helicity for K19 and DArg19, bearing a positively charged side chain at the C-terminus.

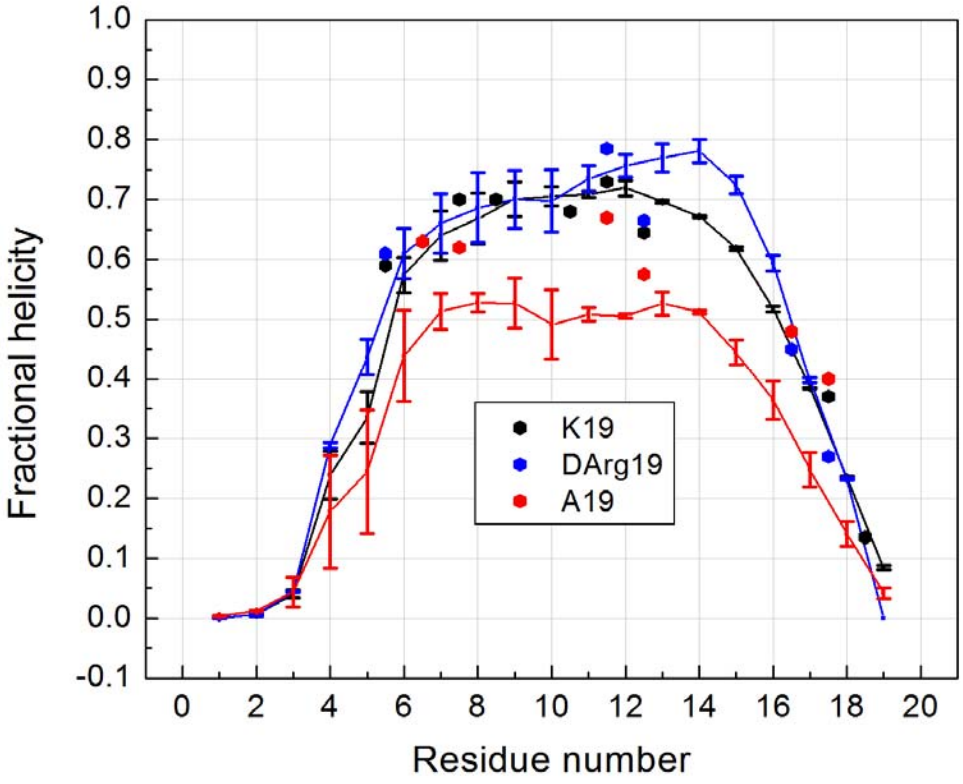

Figure 2-3. Fractional helicities at each residue of peptides A19, K19 and DArg19. The continuous lines connect per-residue values obtained from REMD ensembles at 275 K. The experimental values (points) are from ^{13}C-NMR for the corresponding (Ac)YGG-capped peptides; the A19 data was for the peptide lacking the N-terminal acetyl. Since the NMR data reports on the amide linkage between residue i and i+1, we place the experimental point based on the ^{13}C=O CSD of residue i halfway between i and i+1 on this plot.

The fractional helicities at each residue generated from NMR data and from the simulation ensembles at 275K are compared in Figure 2-3. Both methods show that the

structure ensembles have sequence-dependent helicity measures, and the results are in good agreement, suggesting that further detailed analysis of the simulated structure ensemble is warranted. The relatively high helical content and flat profile across these sequences (Figure 2-3) suggest that the most populated conformation for the peptides is a single helix, rather than isolated smaller helical fragments. A single helix was also demonstrated to be preferred in Nymeyer and Garcia's explicit solvent REMD simulation of the peptide Fs-21(*94*). However, Nymeyer and Garcia's simulations using GB solvation indicated that a single helix structure was not favored. This is in disagreement with our present results using GB, but the differences may arise from differences in protein force field model employed in the two studies or possibly sensitivity of the GB model to the identity of the charged residue (D-Arg vs. Lys). The effects of the C-terminal residue are somewhat greater in the simulations than in the experimental data. The increasing helicity for the C-capping series, Ala $\rightarrow$ Lys $\rightarrow$ D-Arg is experimentally verified for the (KAAAA)3-XGY-NH2 peptides. Since the helical profiles from simulation and experiment are in reasonable agreement at low temperature, we examined the stability as a function of increasing temperature (Figure 2-4). The melting curves for seven Ala residues in an XGG-$(KAAAA)_3$K-span were obtained through both NMR experiments and simulations. These residues were chosen to probe the termini and the central portion of the sequence (which would be affected by equilibrium between single helix and helix-turn-helix conformations). Melting curves for the simulated ensembles were calculated in two ways: from local phi/psi values (panel A) and from the DSSP algorithm (panel B).

The fractional helicities calculated from the chemical shift melts (inset in Figure 2-4 A) are in closer agreement with those from Lifson-Roig theory based on the phi/psi angle. The melting curves for the central and N-terminal repeat sites were very similar, with enhanced melting observed in the less structured C-terminal repeat, particularly in the experimental data. In general, melting curves generated by the simulations are flatter than the ones from the NMR experiments, likely due to lack of temperature dependence in the continuum solvent model that we employed.

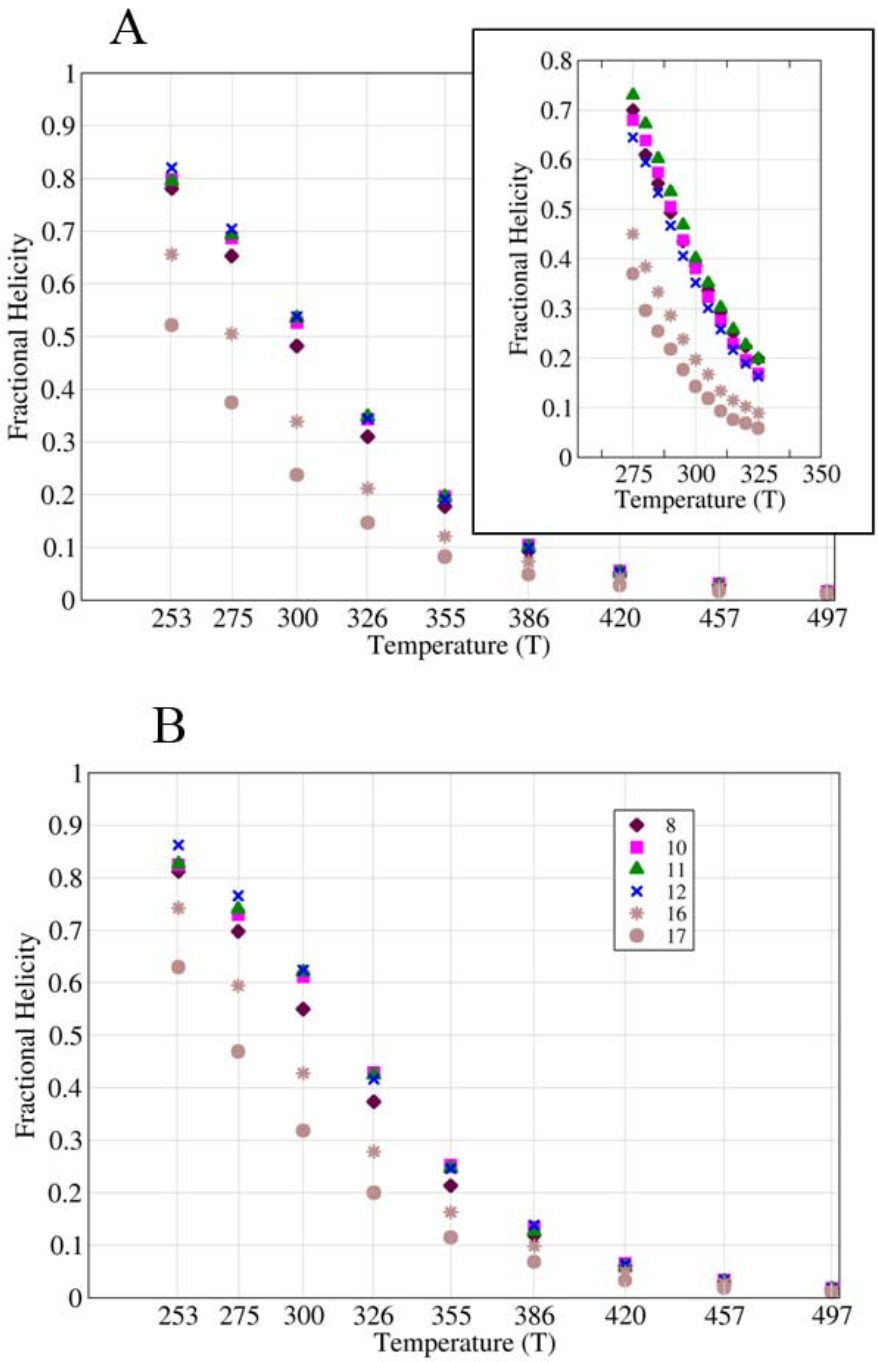

Figure 2-4. Average α-helical propensities of representative residues at different temperatures. Different symbols are used for each residue examined.

Helical propensity was calculated using local backbone conformation of the residue (left) and DSSP (right). The inset in A shows the NMR shift melts.

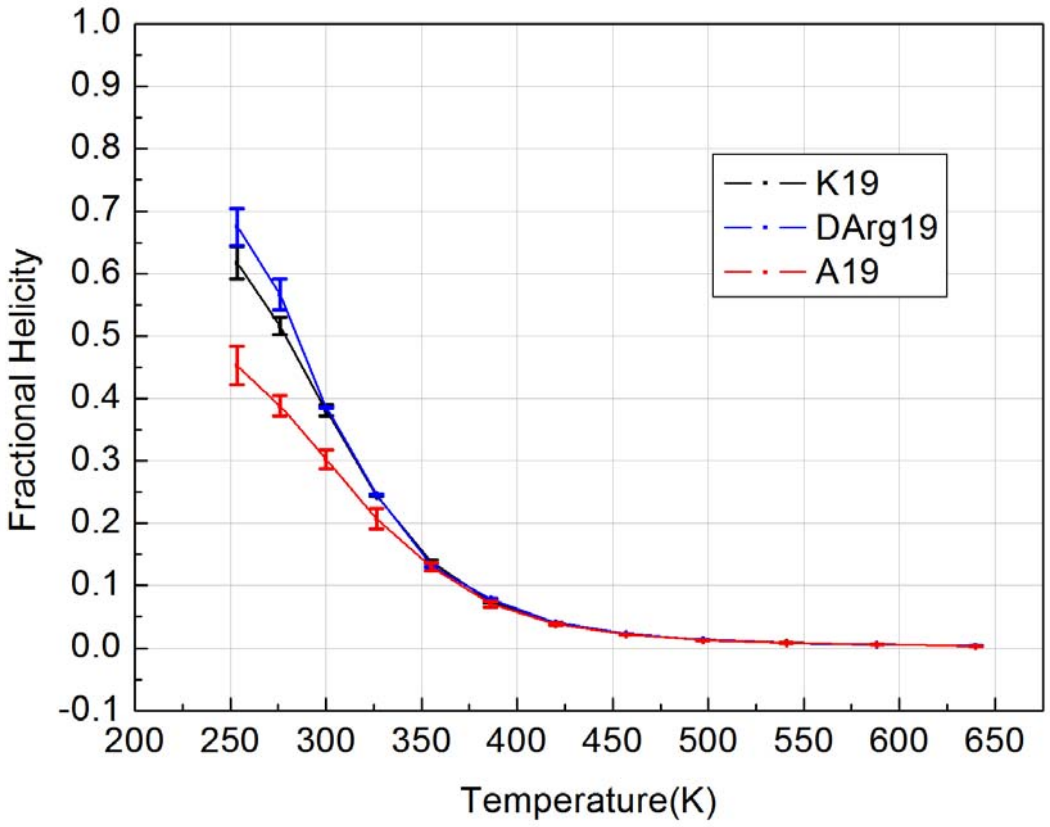

Figure 2-5. Lifson-Roig based melting curves for the helical segment (K4 through X19, normalized for this region) of the three peptides.

Figure 2-5 shows the melting profile for average overall helical content in the peptide as assigned by Lifson-Roig theory. Two interesting observations can be drawn.

One is that the D-Arg C-capped sequence displays enhanced stability at the lower temperatures. One possible rationale is that only a particular D-Arg conformation can stabilize the helix structure; as D-Arg becomes more disordered, the Lys and D-Arg capped sequences have similar helicities at temperatures higher than 300 K.

2.3.3 Number and length of helical segments

As previously noted, the high α-helical content at 275 K and flat profile across the central portion of the sequence (Figure 2-3) suggest significant population of a conformation consisting of a single α-helix. To test this hypothesis, we calculated the population of structures in the ensemble containing different number of continuous α-helix segments using the Lifson-Roig model, in which a helical segment is defined as 3 or more continuous residues adopting helical backbone conformation. We note that the resulting helix conformation populations (Figure 2-6) only take into account the number of helices, not their length (which will be discussed below). At 275 K, more than half of structures (64(±0.5)% for A19, 69(±1)% for K19, and 71(±1.5)% for DArg19) contain only a single continuous helix; this is consistent with our hypothesis based on Figure 2-3, the results reported by Nymeyer and Garcia for explicit solvent simulations(*94*), and the general practice of applying the single sequence approximation in helix/coil treatments of short peptides. The population of conformations with a single helix decreases with increasing temperature slightly faster than the two helix conformation populations.

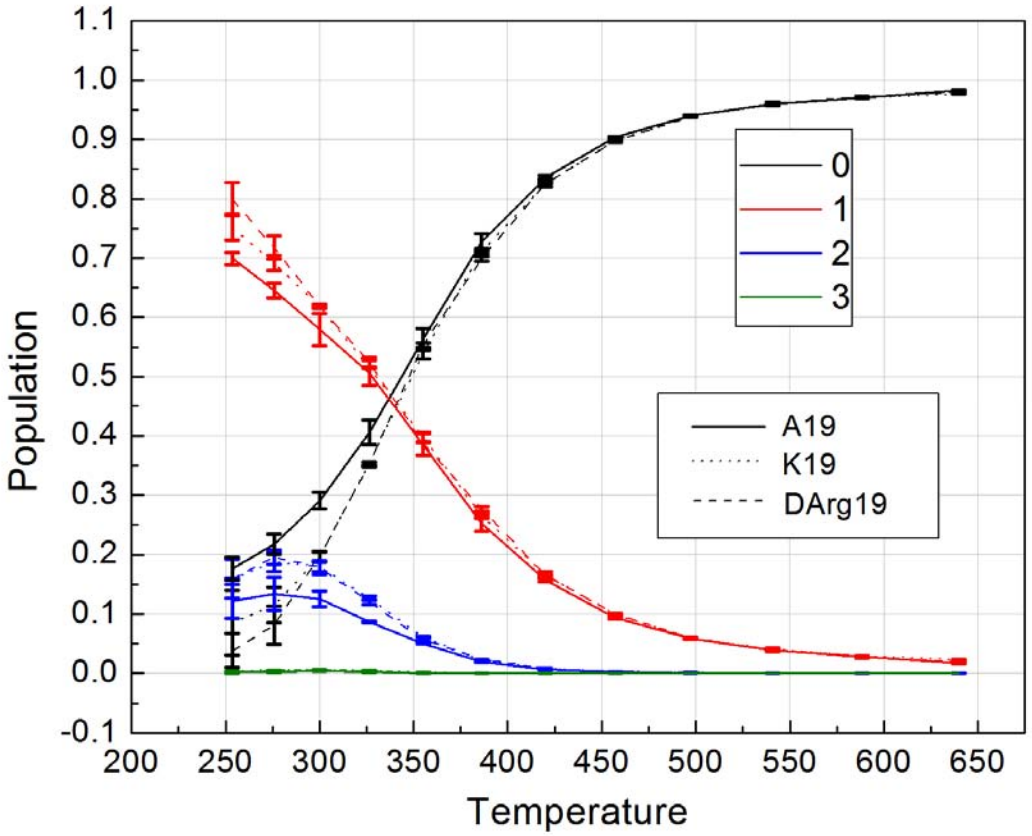

Figure 2-6. Temperature-dependent population of structures containing different number of continuous helical segments sampled in the REMD simulation ensembles. The numbers in the legend refer to the number of continuous helical segments. Different peptide sequences are indicated using different line styles.

At low temperatures (below 325K), the only other conformations with significant population are those containing two α-helical segments, which includes helical bundles (helix-turn-helix motifs). This type of structure has been observed in previously reported

studies of the Fs-21 peptide using REMD simulation(*94*) and normal MD simulation(*93*). Similar to the single helix structures, helical bundle conformations are most populated at low temperatures (18% population at 275 K), and melt with increasing temperature. Structures containing three or more helices are insignificantly populated (<0.6%), consistent with a cooperative effect in helical formation in this sequence(*111*).

2.3.4 Free energy landscape and structure families

To illustrate the structural ensembles sampled in our simulation, we constructed the free energy landscape using principle component analysis (PCA) of the ensemble sampled at 275K. Previous analyses showed that the three peptides have similar overall structural ensembles, therefore the peptide K19 was chosen as a representative for PCA landscape analysis. As described in Methods, the largest eigenvectors from PCA were used to define reaction coordinates for the landscapes, and relative free energies along these coordinates were calculated from histogram populations. These two eigenvectors represent only 42% of the total fluctuations in the system, suggesting that more than 2 coordinates are required for a comprehensive view of the free energy landscape even for this short peptide. Since the PCA eigenvectors, and therefore the local minima, do not directly provide a structural interpretation, we complemented these landscapes with cluster analysis of the ensemble of structures. Following cluster analysis, eigenvalues were calculated for samples of structures in each of the six most populated clusters (which together represent 78% of the ensemble), and these were projected onto the

landscape in order to clarify the structural properties of the different basins. The results are shown in Figure 2-7, along with representative conformations of these six clusters. The arrows indicate their positions on the free energy landscape.

The representative structures from the clusters suggest the general features that may be adopted by structures sampling those basins. In order to more fully characterize the extent of secondary structure formation in these families, we calculated average helical content for each residue, similar to Figure 2-3 except that each cluster was considered separately. The results are shown in Figure 2-8.

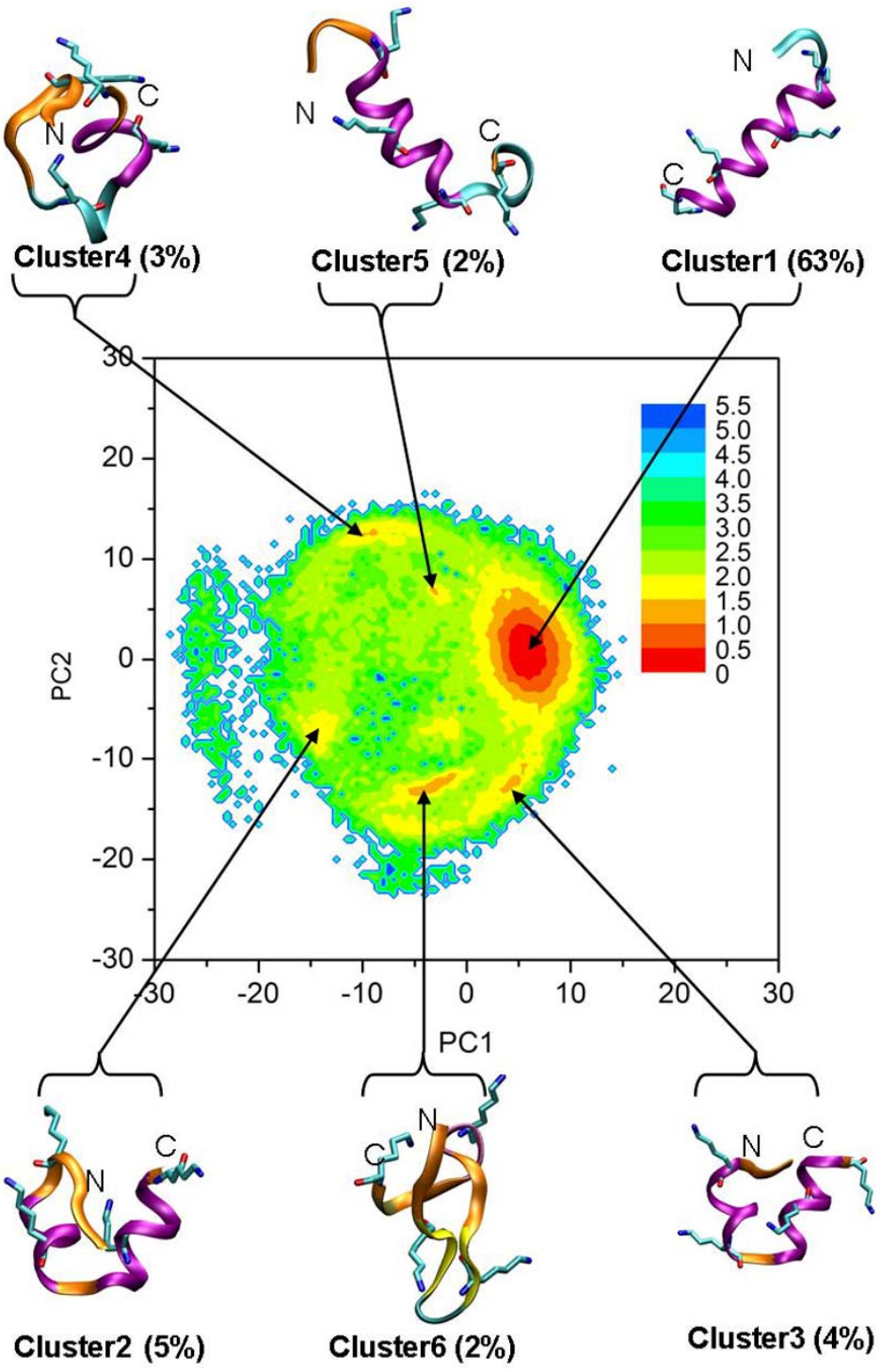

Figure 2-7. The free energy landscape for K19 at 275 K, along with representative structures obtained from cluster analysis of the ensemble. X and

Y axes represent the first 2 principle components. Relative free energy values (in kcal/mol) are represented by color as indicated by the legend. Representative structures for each basin are shown. The colors of the structures reflect secondary structure type: alpha helix – purple, extended beta – yellow, turn – cyan, coil – orange. Only lysine side chains are shown. The populations of the clusters are shown in parentheses.

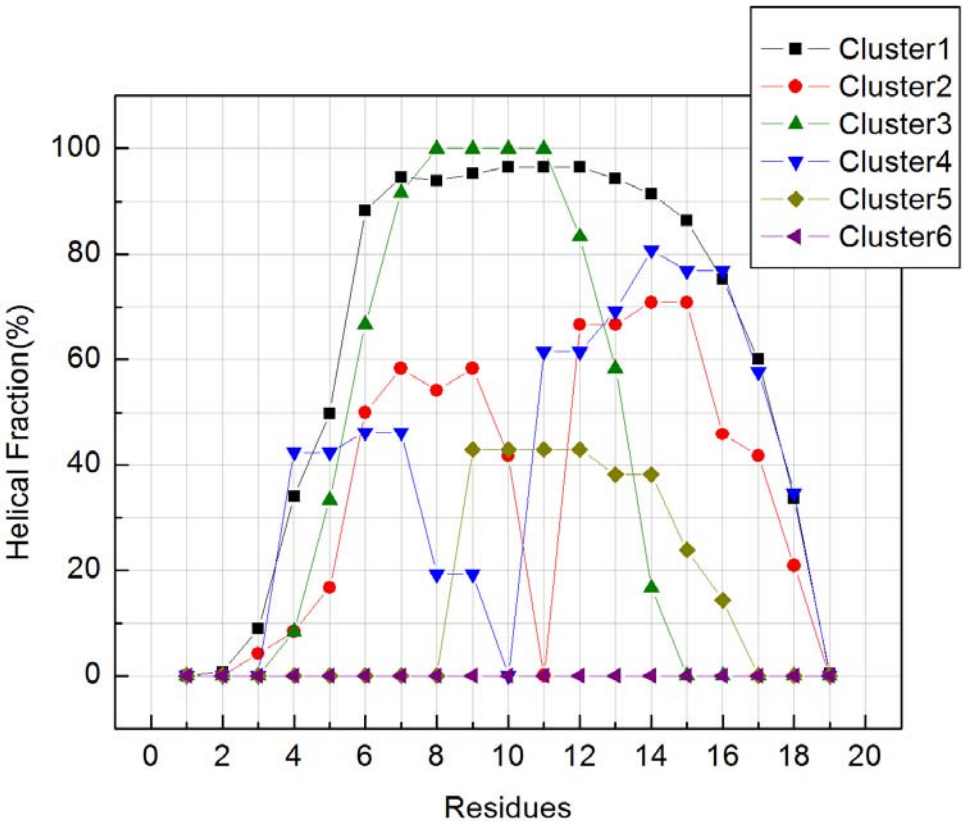

Figure 2-8. The normalized average helical propensity as a function of sequence for each cluster shown in Figure 2-7. Similar to the overall ensemble shown in Figure 2-3, helical content is reduced at the termini of most clusters.

Most clusters also have nearly flat profiles along the middle of the sequence, with the exception of clusters 2 and 4 which correspond to helix-turn-helix motifs. Cluster 6 shows no helical content.

This free energy landscape demonstrates that this peptide has a well-defined and deep global free energy minimum at 275K, approximately 1.5-2.0 kcal/mol lower than alternate local minima. Consistent with the Lifson-Roig analysis, the global free energy minimum corresponds to a family of conformations with a single long helix (Figure 2-7) with significant helix for residues 5-17 (Figure 2-8). The population of this cluster is ~63%, much higher than the next most populated clusters (only 4-5%). Other than the clusters adopting a single long α-helix, the two most populated clusters (cluster 2 and 4, with 4-5% population in each) are both α-helix bundle structures, with each corresponding to a different local minimum on the free energy landscape. Both have turns near the middle of the sequence, as shown in Figure 2-8, but the turns occur at different points. This diversity in the location of the turn or coil in structures with 2 helical segments is also reflected in the observation that structures with 2 helical segments comprise ~20% of the low-temperature ensemble (Figure 2-6), yet none of the individual conformation clusters have a population of more than 5% .

Figure 2-8 provides insight into basis for the end-fraying seen in Figure 2-3. The diminished helicity in the early portion of the N-terminal repeat represents the contribution of two single helix conformers that begin later in the sequence and the lesser

helicity in the N-terminal vs C-terminal helices of the two helix bundle conformations. All clusters within the ensemble display a similar progression of helix fraying at the C-terminus.

In order to investigate the specific interactions that determine the relative stabilities of these conformation types, we performed energy and hydrogen bond analysis on the structures in each of the six largest clusters. Average values for number of intramolecular hydrogen bonds and energy components are shown in Table 3-2. The single long helix cluster (cluster 1) has an average of 2-3 more intramolecular hydrogen bonds than the other conformations. The energy components are also significantly different from the other clusters. Cluster 1 has much lower electrostatic energy (>18 kcal/mol lower than the other clusters) arising from formation of the (i, i+4) backbone hydrogen bonds, and relatively higher solvation energy (EGB) (>10 kcal/mol higher than other clusters) due to the desolvation of the amide groups accompanying helix formation. The other clusters show similar energies, except for the most weakly populated cluster 6, which has no helical content.

	# Intramolecular H-bonds	Electrostatic	VDW	EGB	Sum	Free energy
Cluster1	11.3	142.8	-29.4	-488.3	-374.8	0
Cluster2	8.3	168.3	-33.4	-504.9	-369.9	1.43
Cluster3	9.1	160.7	-30.9	-498.8	-369.1	1.45
Cluster4	8.8	167.2	-31.2	-504.7	-368.8	1.70
Cluster5	8.3	162.5	-26.4	-503.7	-367.6	1.78
Cluster6	6.4	170.6	-32.3	-503.6	-365.3	1.83

Table 2-1. The average number of hydrogen bonds and average values of nonbonded energy components (in kcal/mol) for the six structural clusters shown in Figure 2-7. Electrostatic is the intramolecular electrostatic energy. VDW is the intramolecular van der Waals energy. EGB is the electrostatic component of solvation energy as calculated by the generalized Born model. The free energy relative to cluster #1 is calculated based on the population of the clusters. The total energy is lowest for cluster 1, which has the highest population, and similar for all of the other clusters except #6, which has no helical content. Analysis of the components is provided in the text.

Similar helix-turn-helix structures have been reported in other simulations of other short helix-forming peptides. In their replica exchange simulations of Fs-21(*94*), Nymeyer and Garcia showed that the structures that were most populated at 200K when using the same GB solvation model and radii set that we employed were helical bundles,

with fewer than 9% population of single helix structures. This was in contrast to their results with an explicit solvent model, in which a single helix was preferred. They suggested that the inconsistency was due to inaccuracies in the GB model. Our results (albeit on a different helix-forming sequence) are in disagreement; the GB solvent model does result in significant population (62%) of single-helix structures at 275 K, more than 10 times the size of any other conformation family. Two other helix formation simulation studies with GB solvent models(*93*, *95*) also showed that the single-helix structures were the most populated at temperatures below 300 K. It is possible that the relatively short length of Nymeyer and Garcia's simulations (7ns) resulted in poor convergence of populations. A recent replica exchange simulation study on a helix-forming peptide showed that 15 ns simulation was required for convergence(*95*). Our simulations were 56 ns long, with convergence of the population of cluster 1 only reached after 10ns (data not shown).

Another possible reason for the discrepancy could be differences in backbone parameters employed in the studies, although the high population of single helix conformation that they observed using explicit solvent suggests that this is not the cause if the explicit solvent data is well converged. Another possible reason for the discrepancy is the sequence difference of the two peptides. Peptide K19 has solubilizing Lys residues rather than the Arg residues present in the Fs peptide. It has been shown that the GB model performs more poorly for Arg than Lys, leading to insufficient desolvation penalty for the Arg guanidinium group upon formation of intramolecular hydrogen bonds(*46*).

Helix formation in Fs-21 was also studied by standard MD simulations with a GB

solvent model(*93*). In their studies, Zhang *et al*. used 25 100-ns standard MD simulations to sample the conformations at 273 K and 300 K. They found that the single-helix structures are the most stable conformation at 273 K and the helical bundle structures are the most stable conformation at 300 K. According to their calculations, the helical fractions of the residue 10 are significantly lower than the neighboring residues (50% lower at 300 K and 30% lower at 273 K), corresponding to the formation of helical bundle states. This phenomenon is not reproduced in our simulations or in the experimental measures of fractional helicity. Our experimental data and simulation studies both show that the residues in the middle of the sequence display similar, maximal fractional helicities: the populations of helical bundle structures must always be much lower than the populations of single helix conformations. The difference between Zhang *et al.*'s and our computational ensembles could result from the improved sampling and convergence obtained in our studies using REMD, differences between force fields used in that study and the present one, or simply different conformational preferences of the peptides studied.

2.3.5 The radius of gyration at different temperatures

The radius of gyration is used to characterize the effective size and shape of the molecule. Figure 2-9 shows the radius of gyration (R_{gyr}) distributions at twelve temperatures. To clarify the representation, the data are divided into groups according to temperature. Four temperatures, 253 K, 276 K, 300 K and 326 K are shown in Figure 2-9A. At these four temperatures, there are two peaks in the R_{gyr} distribution. One is located at 7.5 Å, which represents the single helix structures (cluster 1 in Figure 2-7). The

other peak is near 6 Å, which represents more compact structures such as partial helices, helical bundles and coils. The 7.5Å peak is most populated at temperatures below 300 K. With increasing temperatures, the height of the helix peak is reduced, and the height and the width of the globular state peak are increased slightly. At 354 K (Figure 2-9B), the peak for the single helix structure disappears as it melts; only the globular and partial helix conformations remain. With increasing temperature, the peak shifts to higher values. At temperatures above 450K, the ensemble is composed of random coil structures, with a broad Gaussian distribution of Rgyr centered about 9.5 Å.

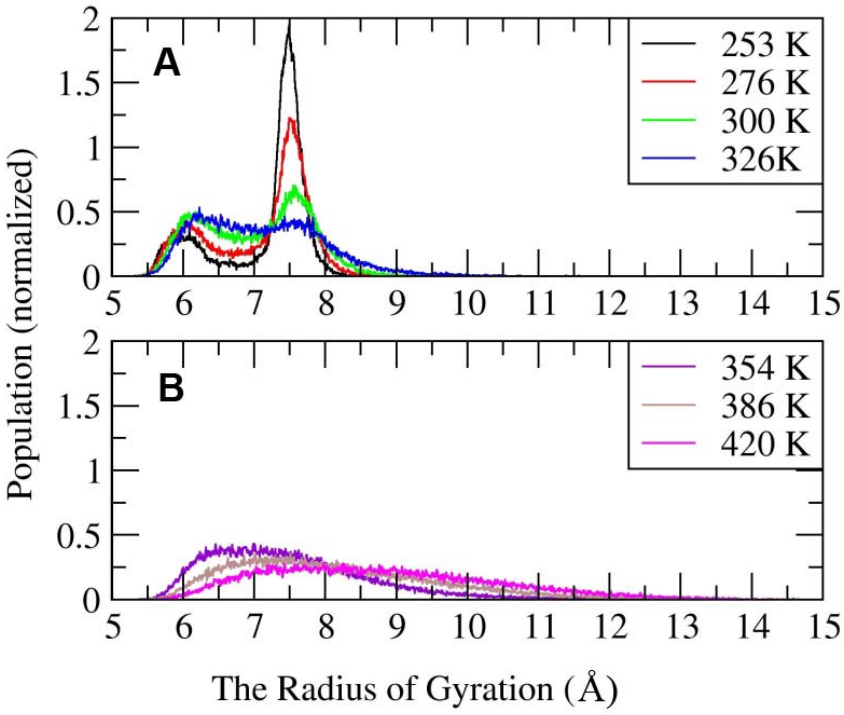

Figure 2-9. The distributions of the radius of gyration at different temperatures.

2.3.6 Role of the lysine sidechains

Helix simulations(*67, 96*) and some experimental studies(*92, 112*) have shown that amino acids with charged sidechains can significantly enhance α-helix formation. Other experimental data indicate that central lysines in helices have a net destabilization effect(*97*). The present study confirms the helix-favoring effect of a C-terminal Lys, which likely reflects a Coulombic interaction with the helix macrodipole rather than backbone desolvation or an H-bonding interaction. Simulation data reported for A21 (Ac-Ala21-methyl amide), which lacks a favorable Coulombic effect at the C-terminus, suggest that it is not well structured (<40% average helical content at 275 K)(*67*). Some simulations indicate that the stabilization of helical content by the charged side chains arises from shielding of the intramolecular backbone hydrogen bonds from the solvent in an α-helix conformation. Previous simulations using GB indicated that helical conformations were actually destabilized by the presence of ionizable sidechains, due to competition between α-helical i-i+4 backbone hydrogen bonds and those involving the arginine guanidinium groups and backbone carbonyl oxygen atoms(*94*). These backbone-sidechain interactions were not observed in similar simulations using explicit water molecules.

These continuing issues prompted as to examine the role of lysine sidechains in our simulations. To determine the extent of Lys sidechain interaction with the backbone, we calculated hydrogen bond populations in the simulated ensembles of peptide K19. Lys9 was selected as a representative example because its location in the central region of the sequence would allow interaction with backbone residues on either the N-

terminal or C-terminal side. The data shows that no significant sidechain hydrogen bonding is present in the ensemble, with all possible Lys9 Nζ - backbone carbonyl hydrogen bonds having less than 1.5% population. Other residues are similar; hydrogen bonds involving the sidechain amino group of Lys19 have populations less than 5%.

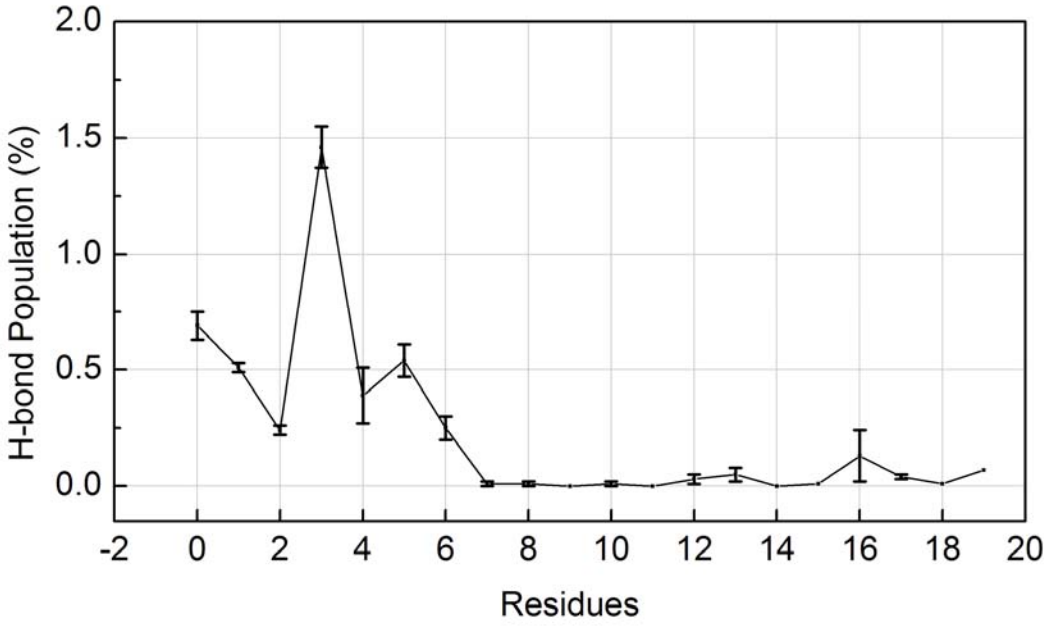

Figure 2-10. The population of hydrogen bonds formed between the Lys9 sidechain amino group and the backbone carbonyl oxygen of each residue. The x axis is the acceptor residue number. The y axis is the percentage occupancy of the hydrogen bond in the ensemble. Residue 0 corresponds to the N-terminal acetyl group

To investigate lysine's potential α-helix stabilizing effects, we calculated the correlation between helix formation and Lys9's sidechain position. To quantify the position of the sidechain, two distances were measured: one is the distance between Ala5 Cα and Lys9 Nζ, the other is the distance between Ala13 Cα and Lys9 Nζ. The preference of Lys to interact with the N- and C-terminal end of the chain is indicated by the difference in the two distances. At low radius of gyration values corresponding to compact coil and helical bundle conformations, a broad, flat distribution of Lys distances is observed with no apparent preference for N-terminal or C-terminal direction. A slight bias toward the N-terminus may reflect the formation of partial helices in that region. For larger radius of gyration values corresponding to the helical conformations, there are two nearly equally populated conformations for the Lys9 sidechain. In one, the amino group is 5 Å closer to the Ala5 than to Ala13. In the other basin the two distances are similar.

To verify this interpretation, the structures belonging to the single helix cluster were re-clustered using the Lys9 side chain conformation. As expected, two families were found and the representative structures are shown in Figure 2-12. In one family (Figure 2-12a) the methylene groups of the Lys9 sidechain shields the upstream i-4 residue's backbone carbonyl oxygen from solvent. In the other conformation (Figure 2-12b) the two distances are same; the sidechain is perpendicular to the axis of the helix. Interestingly, two highly similar conformations were observed in previously reported explicit solvent simulations of the Fs peptide (*67*). Even without explicit water molecules, our simulations exhibit the same feature of protection of backbone hydrogen bonds by the nonpolar region of the ionizable sidechain, with the charged group interacting with

solvent.

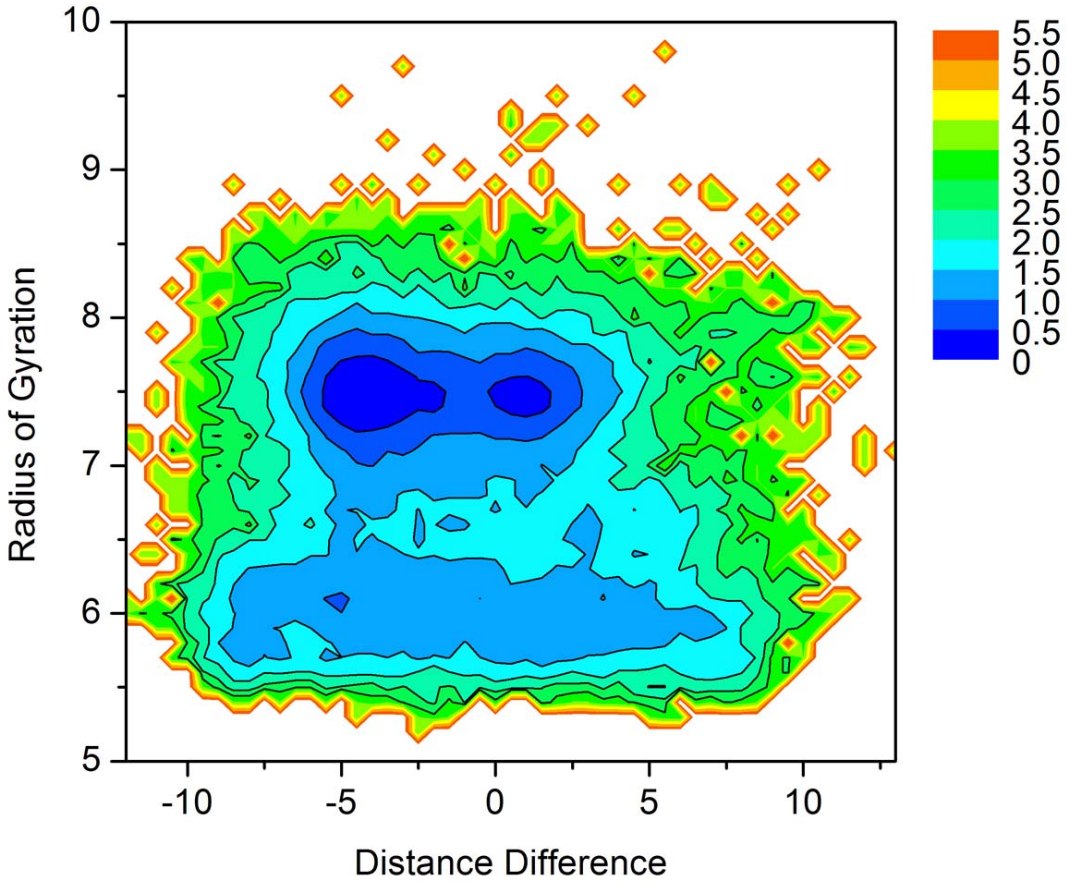

Figure 2-11. Free energy surface at 275K indicating conformational preferences for the Lys9 side chain. The X axis corresponds to the difference between the distance from Lys9 Nζ to Ala5 Cα and to Ala13 Cα. Values near zero indicate no preference, while positive and negative values indicate a shift toward the C- or N-terminal end of the helix, respectively. Color indicates relative free energy; values are given in kcal/mol.

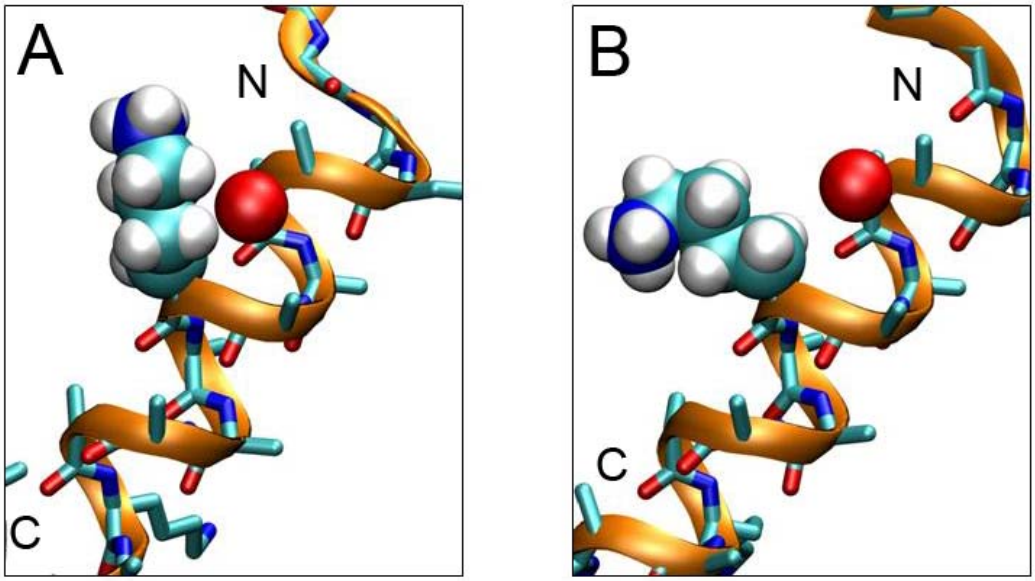

Figure 2-12. Two representative conformations of the Lys9 sidechain in single helix structures of 3Ai. Only backbone and Lys9 sidechain atoms are shown.

Confidence in the conformations of Lys's sidechain we observe in simulations is strengthened by the NMR experiments done by Groebake *et al.*(*113*). They have used ^{1}H NMR spectroscopy to illustrate the conformation of the Lys's sidechain in α-helix. Medium intensity NOE interactions were observed between the (i-3) Ala α-H and β-, γ-, δ-H protons of the Lys, and a weak interaction with the ε-H. In the random coil conformations, these NOE interactions were absent. To compare with these NOE results, we calculate the distances between the α-CH of Ala6 and the β-, γ-, δ- and ε-H of Lys9 in

the single helix cluster (generated by cluster analysis) and in the random coil cluster (the structures with no w-state residue). The values are converted into ensemble average by equation: $\sqrt[-6]{\langle dist^{-6} \rangle}$. The results are shown in Table 2-2. In the single helix cluster, the ensemble averaged distances are all between 3.0 and 4.0 Å. Furthermore, ε-H, which has a weak interaction with Ala6's α-H in Groebake's experiment, has the longest distance in our simulation as well. In the random coil cluster in which no NOE signal has been observed in Groebake's experiment, all distances are longer than 4.0 Å. Taking into account the differences of sequence and methology between our simulation and Groebake's NMR experiment, it can be concluded that our results are qualitatively consistent with the NMR experiment.

	Single helix cluster	Random coil cluster
β-H	3.0	4.2
γ-H	3.9	4.6
δ-H	3.3	4.6
ε-H	4.0	4.8

Table 2-2. The ensemble averages of distances between Ala6's α-H and the protons in Lys9's sidechain.

2.4 Conclusion

We have studied the temperature dependence of helical propensities for the peptides Ac-GGG-$(KAAAA)_3$X-NH2 (X = A, K, and D-Arg) using replica exchange molecular dynamics simulations and the generalized Born continuum solvent model. The simulation results are compared with data obtained from NMR chemical shifts of -$GG(KAAAA)_3X$-NH_2 and Ac-$(KAAAA)_3XGY$-NH_2 sequences (X = A, K, and D-Arg) and good agreement is found with both the absolute helical propensities as well as relative helical content along the sequence. The temperature dependence is also in reasonable agreement with the experimental results. Thermodynamic parameters calculated based on the melting curve from Lifson-Roig analysis of the simulation ensembles are comparable with experimental results on a related sequence. Based on this data, detailed structural analysis of the simulation ensembles was performed. Cluster analysis showed that the global minimum on the calculated free energy landscape corresponds to a nearly fully α-helical conformation. Helical bundle conformations were populated, but these local minima were 1.5-2.0 kcal/mol higher in free energy and were less populated at all temperatures simulated. Energy component analysis shows that the single helix state has favorable intramolecular electrostatic energy due to hydrogen bonds, and the globular states have favorable solvation energy due to the exposure of the polar atoms. While both experimental and simulation studies shown increasing helicity in the series X = Ala $\rightarrow$ Lys $\rightarrow$ D-Arg, none of the data sets show a specific D-Arg effect as large as that quoted (1.2 kcal/mol helix stabilization) in the initial reports on this C-capping function(*98*). An increase in helicity due to placing either Lys or D-Arg at the C-terminal position can be

rationalized as an interaction with the helix macrodipole. The analysis of lysine sidechain conformations suggests that polar residues can favor helix formation by protecting the backbone hydrogen bonds. A structural analysis of C-terminal ending group D-Arg in the simulated helical ensembles shows that it can from H-bonds with Ala16, by doing so to slightly increase the stability of helical structures. A resolution of the questions concerning the thermodynamic effects of C-capping by positively charged sidechains, particularly of the D-Arg function, will require studies of additional peptides with both L- and D-configured terminal groups and other polar functions. The agreement in helicity profiles for these sequences in experimental and molecular dynamics ensembles indicates that MD will continue to be a useful method for examining the details of polypeptide secondary structuring.

Chapter 3 Computational analysis of the binding mode of 8-oxo-guanine to formamidopyrimidine-DNA glycosylase

3.1 Introduction

DNA is a major target of oxidative damage which, in turn, has been linked to human diseases associated with aging, including cancer (*114, 115*). 8-oxoguanine (8OG) (*116*) is one of the most common forms of oxidative DNA damage(*11*); failure to repair this lesion prior to DNA replication leads to G:C to A:T transversion mutations in bacterial and mammalian cells(*22*).

In *E. coli*, Fpg (MutM), MutY and MutT work in concert to counter the potentially deleterious effects of 8OG (*117, 118*). Fpg is an 8-oxoguanine-DNA glycosylase/AP lyase which excises 8OG from oxidatively damaged DNA. This multi-step process is initiated by nucleophilic attack on C1' by the N-terminal proline of Fpg, forming a Schiff base intermediate. Proton abstraction leads to β-elimination of the 3' phosphate, followed by hydrolysis of the Schiff base and δ-elimination, generating a single base gap (*10, 21*).

While the catalytic mechanism by which 8OG is excised from DNA has been extensively investigated by structural methods(*25, 27, 119-123*), little is

known regarding the mechanism by which Fpg recognizes its cognate lesion and whether discrimination between the oxidized base and guanine occurs during one or several stages of binding. 8OG differs from guanine at the N7 and O8 positions in that N7 is protonated and the C8 hydrogen is replaced by oxygen (Figure 1-4). It seems likely that that these structural differences account, at least in part, for the ability of Fpg to recognize and bind 8OG in its active site.

The initial challenge in determining structural requirements for binding specificity through X-ray crystallographic analysis was presented by the difficulty in trapping the Fpg-DNA complex before base excision occurs. Two approaches have been used to circumvent this problem. Gilboa *et al*. used sodium borohydride to trap the Schiff base intermediate by forming a covalently- bound intermediate (*25*). The resulting crystal structure of the *E. coli* Fpg-DNA complex lacked the damaged base; moreover, the βFα10 loop (residues 217-224) does not appear in the resulting electron density map. Thus, direct evidence for substrate recognition was lacking.

More recently, Fromme and Verdine reported the crystal structure of the Fpg-DNA complex containing the damaged base (*121*) (Figure 3-1). In this case, *Bacillus stearothermophilus* (*B. st.*) Fpg was used for the study with its catalytic activity eliminated by use of an E2Q mutant (*124*). This approach generated a structure with 8OG bound to Fpg in the *syn* conformation, in which a single hydrogen bond between the lesion N7 and Ser219's backbone carbonyl oxygen is formed. The mutation involves replacement of an ionizable (Glu) side chain with a neutral residue (Gln) that was observed to directly interact with the lesion, thus the details of the resulting

structure could differ from those of the active, wild type enzyme.

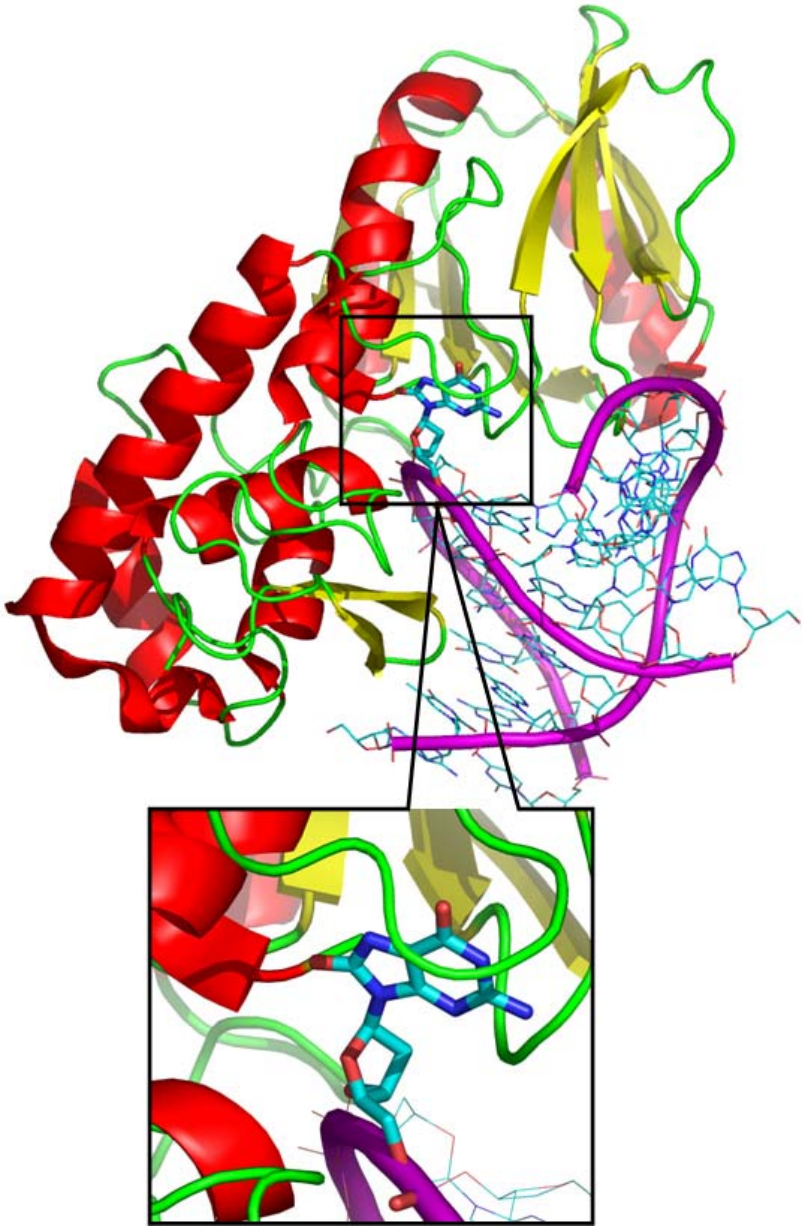

Figure 3-1: Structure of *B. st.* Fpg bound to 8OG DNA (pdb code: 1R2Y). Atomic detail is shown for the DNA duplex and the protein is represented with a cartoon diagram.

Simulation studies (*125, 126*) have been conducted in which the missing base in the structure reported by Gilboa *et al.* was modeled into the crystallographic structure(*25*). Based on their data, the authors proposed a different binding mode for 8OG than that deduced for the E2Q mutant by Fromme and Verdine. These simulations suggested that the 8OG binds to Fpg in the *anti* conformation, making hydrogen bonding contacts to both N7 and O8 of 8OG. In contrast, the crystal structure from Fromme and Verdine show the damaged base to be in the *syn* conformation, with only N7 involved in hydrogen bonding with Fpg(*121*).

Unlike its human analog, hOGG1, Fpg displays no significant π-π interactions with 8OG, resulting in a more spacious binding pocket. Thus, one possible explanation for the apparently contradictory results is that 8OG may bind to Fpg either in *anti* or in *syn*, although this does not fully explain the divergent observations. We previously used a computational approach to study the preference for *anti* and *syn* 8OG conformations in a variety of sequence contexts in duplex DNA(*127*). In the present study, we employ all-atom molecular dynamics simulations and free energy calculations to gain insight into the interactions involved in the complex between 8OG and wild type Fpg. We address two specific questions: what is the preferred conformation of 8OG in the Fpg active site, and what are the interactions between 8OG and Fpg that provide for lesion recognition or a secondary discrimination against guanine prior to base excision? We make the remarkable observation that our simulations are consistent with both models of the Fpg-8OG complex. Additional calculations provide insight into the role of specific

Fpg residues in the binding of DNA containing 8OG.

3.2 Methods

3.2.1 System preparation

All initial structures were built using the Leap module of Amber (version 8) (*128*), based on the crystal structure of the *B. st.* Fpg/DNA complex with 8OG in the *syn* conformation (1R2Y.pdb) (*121*). The sequence of the DNA duplex was d[GTAGACCTGGAC]·[GTCCAG*GTCTAC] (where G* is 8OG). All water molecules in the crystal structure were retained. The *anti* conformation was built by rotating the glycosidic angle of 8OG using the molecular modeling software MOIL-VIEW(*34*). Protein mutants were generated by manual editing of the pdb file, with the new side chain built using Leap. These structures were minimized for 100 cycles of steepest descent and then solvated in truncated octahedron boxes with a minimum 6 Å buffer between the box edge and the nearest protein atom. The TIP3P model(*129*) was used to explicitly represent water molecules. Following previous studies(*125, 126*), the N-terminal proline was modeled as neutral to mimic the stage directly before the reaction. The parameters for neutral N-terminal proline were obtained from Perlow-Poehnelt *et al*. (*126*). Force field parameters for 8OG were obtained from Miller *et al*.(*130*). Zinc was modeled using the Stote non-bonded model (q = +2e-, σ = 1.7 Å, ε = 0.67 kcal/mol)(*131*). The remaining protein and nucleic acid parameters employed Amber ff99 (*100, 132*), with modified protein backbone parameters to reduce the alpha-helical bias of those force

fields(*133*). Previous calculations showed a large pKa shift for glutamate in the active site of T4 Endonuclease V (*134*). However, Fromme and Verdine suggested(*121*) that Glu2 is unlikely to be protonated since it is at the N-terminus of an α-helix; it has been shown that acidic residues in this position tend to have low pKa values (*135*). In this study we calculated the pKa value of Glu2's using two approaches, Jensen's empirical method(*136*) and Onufriev's H++ method(*137*); these methods resulted in Glu2 pKa values of 5.69 and 6.78, respectively. To mimic the crystallization pH (7.5)(*121*) we therefore simulated Glu2 in its deprotonated state unless specified otherwise. We also repeated selected simulations and umbrella sampling calculations using protonated Glu2 to examine the sensitivity of the results to the ionization state.

3.2.2 Molecular Dynamics simulations

All molecular dynamics simulations were carried out with the SANDER module in Amber. Solvated systems were minimized and equilibrated in three steps: (i) 50 ps MD simulation (*128*) with protein and DNA atoms constrained and movement allowed only for water; (ii) five 1000-step cycles of minimization, in which the positional restraints on the protein and DNA were gradually decreased; (iii) Four cycles of 5000 steps MD simulation with decreasing restraints on protein and DNA. A final 5000 steps of MD were performed without restraints. The resulting structures were used in the production runs.

SHAKE(*105*) was used to constrain bonds involving hydrogen atoms. The non-bonded cutoff was 8 Å. The particle mesh Ewald method(*52, 138*) was used to calculate

long-range electrostatics. Constant pressure (1 atm) and temperature (300 K) were maintained by the weak coupling algorithm (*139*).

3.2.3 Structural analysis

The root mean square deviation (RMSD) of three regions of the complex were calculated separately: (i) protein, which is calculated using the Cα atom of all protein residues; (ii) DNA, which includes the backbone heavy atoms of only the 8OG:C and flanking base pairs; (iii) loop, which includes the Cα atom of residues in the base binding loop (222 – 231), with the RMSD calculated for this region after best-fit of the coordinates of the entire protein. The glycosidic angle of 8OG is defined using atoms O4' –C1' –N9–C4. The RMSD and glycosidic angle calculations, along with distance calculations, were carried out using the ptraj module.

3.2.4 Umbrella sampling and potential of mean force calculations

Umbrella sampling (*74-77*) was used to calculate the potential of mean force (PMF) as a function of 8OG glycosidic angle in the binding site. 36 starting structures were generated using MOIL-VIEW(*34*) by rotating 8OG glycosidic angle in 10° increments from 10° to 360°. These initial structures were energy minimized and one independent 200ps simulation (i.e. one umbrella sampling window) was performed for each structure. The glycosidic angle was restrained to the initial value using a harmonic restraint with a force constant of 50 kcal mol-1radian-2. Residues farther than 10 Å from 8OG were restrained using positional restraints (force constant 2 kcal mol-1 Å-2). The other

parameters of these simulations were the same as the standard MD simulations. The resulting PMF was obtained by WHAM analysis(*75-77*) of the data using a program provided by Alan Grossfield (freely available at dasher.wustl.edu/alan).

3.2.5 MM-GBSA method

The MM-PBSA method has become widely used for calculation of free energies of binding(*39*). The MM-GBSA variant, in which the GB solvation model is used, was shown to successfully reproduce relative affinities and selectivities for a range of matrix metalloprotease inhibitors(*140*). In this study, we treated the protein+DNA(without 8OG) as the "receptor" and the 8OG nucleotide as the "ligand". The relative binding energy of 8OG in *anti* or *syn* conformations represents the contribution of protein-8OG interactions to the relative stability of these two binding modes. The absolute binding free energy is calculated by the equation 3.1

$$\Delta G_{binding} = G_{COMP} - G_{receptor} - G_{ligand}$$
$$G_x = E_{VDW} + E_{EEL} + G_{pol} + G_{non-pol} \qquad 3.1$$

The van der Waals energy (EVDW) and intramolecular electrostatic energy (EEEL) were calculated using SANDER(*128, 141*). The polar part of solvation free energy (Gpol) was calculated using the GB-OBC solvent model (*141, 142*). The nonpolar solvation free energy (Gnon-pol) was calculated by using the solvent accessible surface area (SASA) and a surface tension of 5 cal/mol×Å^{-2} (*143*). Energy calculations for the isolated ligand (8OG) and receptor (Fpg) used coordinates sets obtained from the trajectory of the

complex.

3.3 Results and Discussion

3.3.1 Effect of the E2Q mutation

In the crystal structure of *B. st.* E2Q-inactivated Fpg with DNA, the damaged base binds to the active site in the *syn* conformation(*121*) and the βFα10 binding loop is ordered. However, due to the proximity of the Glu2 side chain to the 8OG base (~3 Å), the electrostatic effects arising from the E2Q mutation may be significant. Perlow-Poehnelt *et al.* suggested this mutation as one possible reason for the difference in 8OG conformation in the E2Q *B. st.* crystal structure and the *E. coli* computational model(*126*). We used two sets of simulations to test this possibility and also to investigate the general influence of the mutation on the structure of the complex. One simulation used the same E2Q mutant *B. st.* Fpg-DNA complex that was studied crystallographically and the other employed wild type *B. st.* Fpg in the complex. Three independent 2ns simulations were used for each sequence to evaluate precision.

3.3.2 Stability of the wt and E2Q systems

To analyze the stability of the structures under the simulation conditions, the RMSDs of protein Cα atoms, lesion site residues (8OG:C and flanking base pairs) and the βFα10 loop were computed for each of the six independent simulations. The flexibility of the βFα10 loop was proposed to play an important role in the recognition and excision of the damaged base(*25, 27, 120-122, 125, 126, 144*). Both systems are quite stable in all six

2ns simulations. For the DNA lesion site, the RMSD plateau values are 0.5-1.0 Å with either protein sequence, while the RMSDs of the entire protein and the βFα10 loop region are both slightly larger at ~1.0 Å (Figure 3-2 and Figure 3-3). The loop shows greater fluctuations about this average than the protein or DNA. We thus conclude that the simulations provide stable dynamics and that the mutation does not have any dramatic effect on the overall structure and stability of the complex.

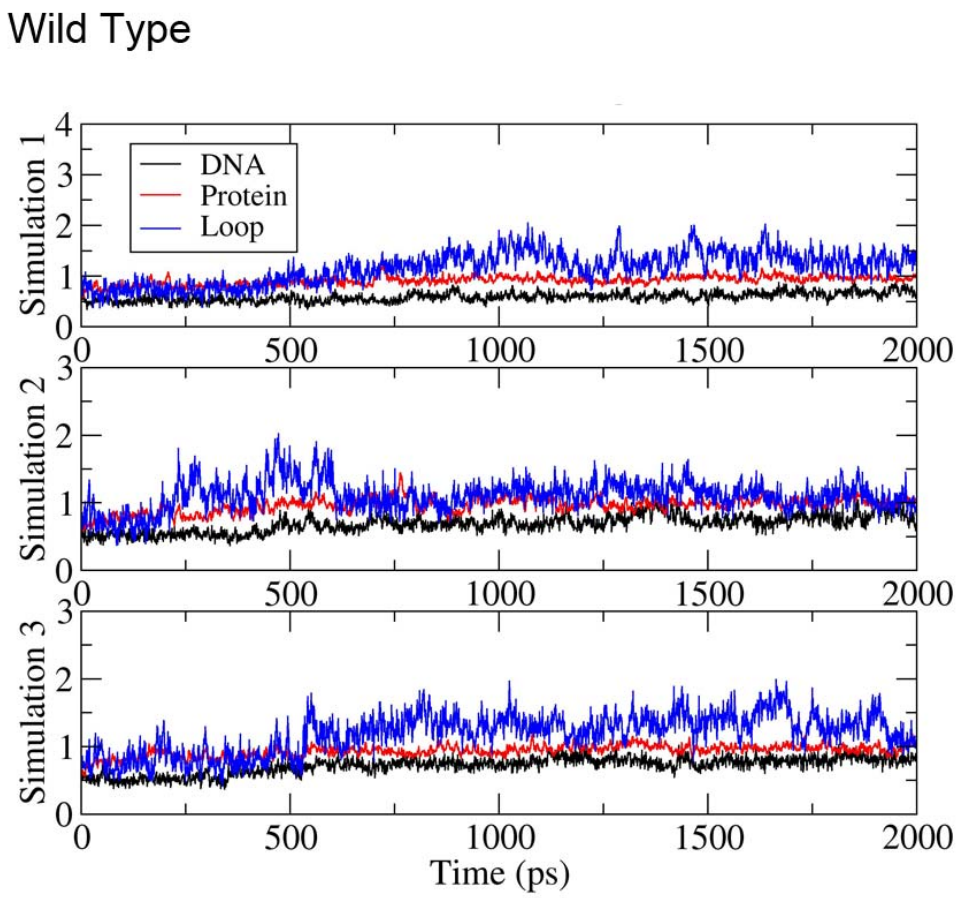

Figure 3-2 The RMSDs of the wild type system. In protein the results are calculated by the Cα atom of all residues in protein. The results for DNA are

calculated based on the backbone heavy atoms of 8OG:C base pair and its upstream and downstream base pairs. Loop inclu7des the Cα atom of residues in the base binding loop (222 – 231), with fitting the reference structure using the coordinates of whole protein.

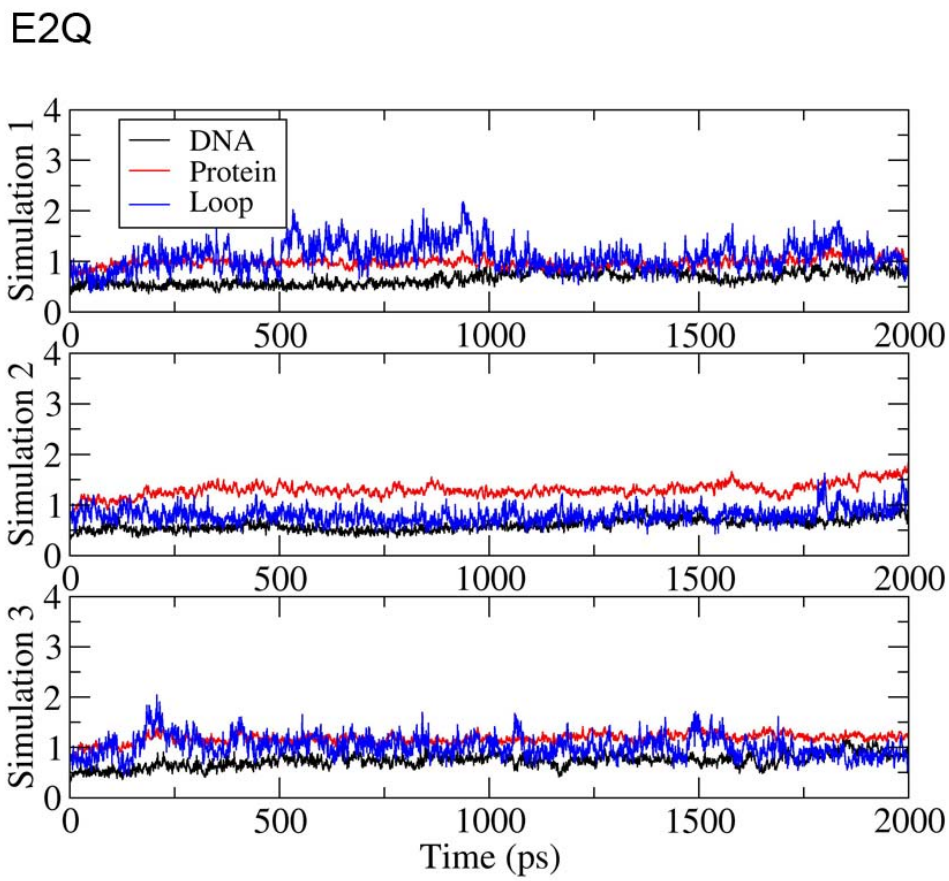

Figure 3-3. The RMSDs of the E2Q mutant system. The legends are defined in the same way as in Figure 3-2.

3.3.3 Influence of the E2Q mutation on the active site geometry

Detailed analysis of the trajectories brought to light several small but interesting differences in the behavior of the two sequences. In the crystal structure of E2Q Fpg, a hydrogen bond forms between the Nε of Gln2 and O8 of 8OG. This interaction is reproduced in the E2Q simulations (Figure 3-4), with an average distance between these atoms of 3.04 Å (Figure 3-6), in excellent agreement with the value of 3.08 Å observed in the crystal structure(*121*). However, this hydrogen bond cannot form in the wild type sequence since the amide nitrogen is not present. In particular, unfavorable electrostatic interactions are present in the wild type between the Glu2 carboxylate and O8 of 8OG, resulting in a shift of the protein strand containing Glu2 away from 8OG in the wt simulations (Figure 3-4).

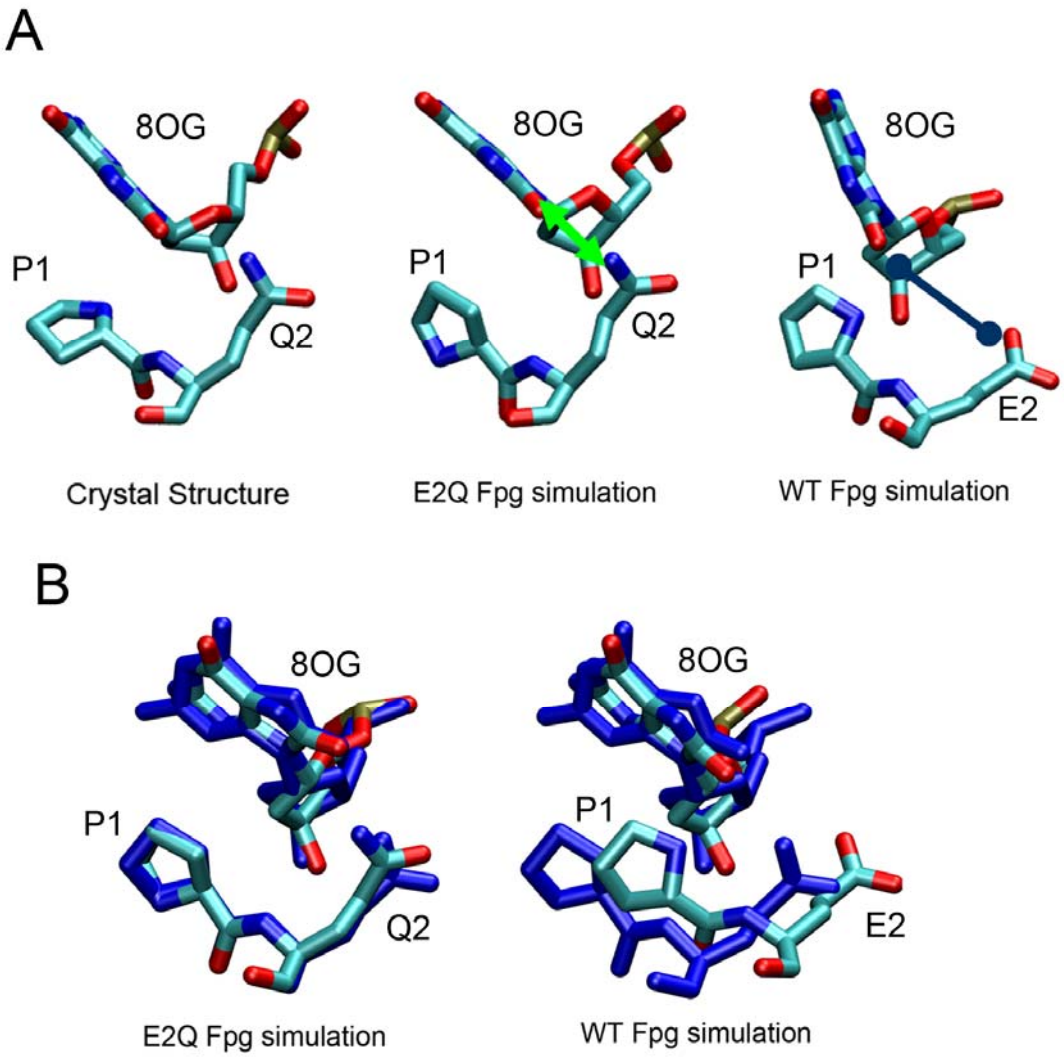

Figure 3-4: Comparison of conformation of the region containing 8OG, Pro1 and Gln2 (Glu2 in wt) between the crystal structure, the E2Q Fpg simulation and the wild type Fpg simulation. The favorable interaction between Q2 and the O8 of 8OG is indicated by a green line and the unfavorable interaction between WT E2 and the O8 of 8OG is indicated with a blue line. Panel B shows the region after best-fit of 8OG, in which the crystal structure is shown in dark blue simulation structures are colored by atom. The E2Q simulations reproduce the E2Q crystal structure, but a shift in P1/E2 relative to 8OG is apparent in the wild

type simulation.

This difference between the E2Q mutant and wt Fpg causes two obvious effects. First, due to repulsion between 8OG and Glu2, the purine ring rotates slightly, increasing the distance between O8 and the Glu2 carboxylate. The glycosidic angle of 8OG in E2Q simulations is 108 ± 21° (uncertainty denotes the standard deviation), similar to 101° in the crystal structure(*121*) suggesting that the simulation is reasonable. In the wild type simulations, however, this angle is significantly different at 57° ± 14°.

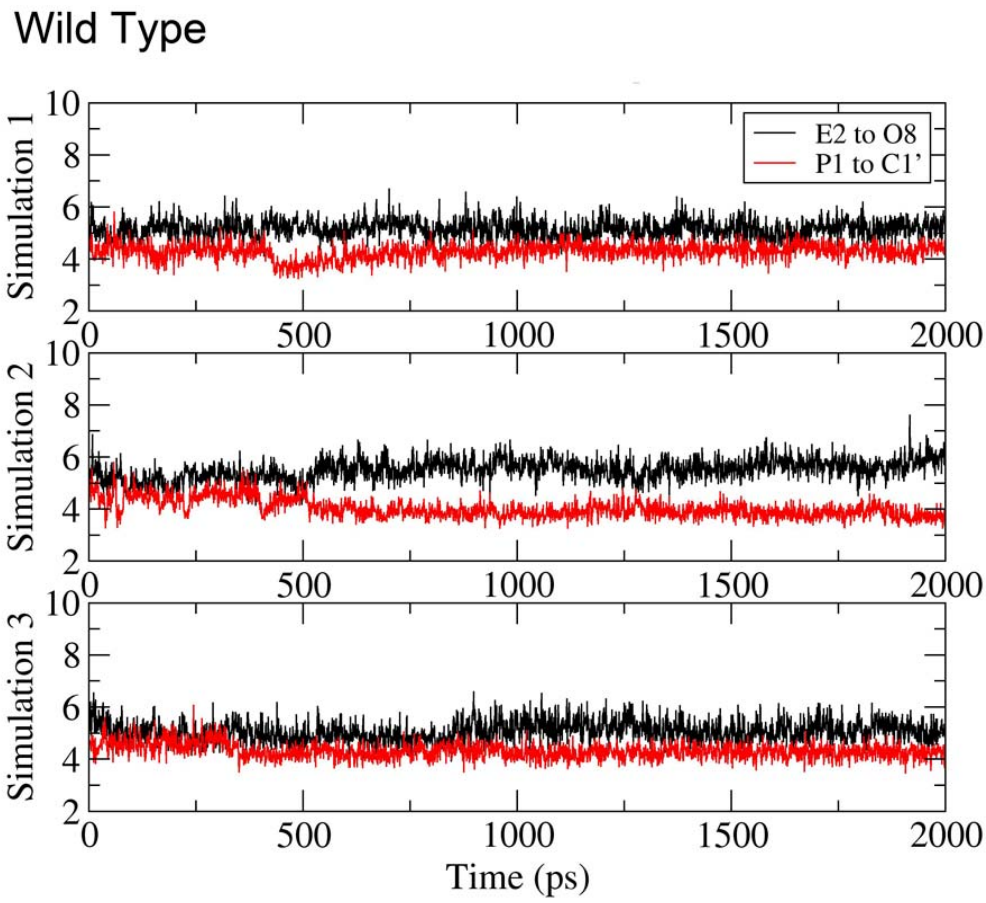

Figure 3-5. The Distance analysis results for the wild type Fpg simulations. The first distance is the one between Pro1's N and 8OG's C1' (labeled as P1 to C1'), and the distance between Oε of E2 and O8 of 8OG (labeled as E2 to O8).

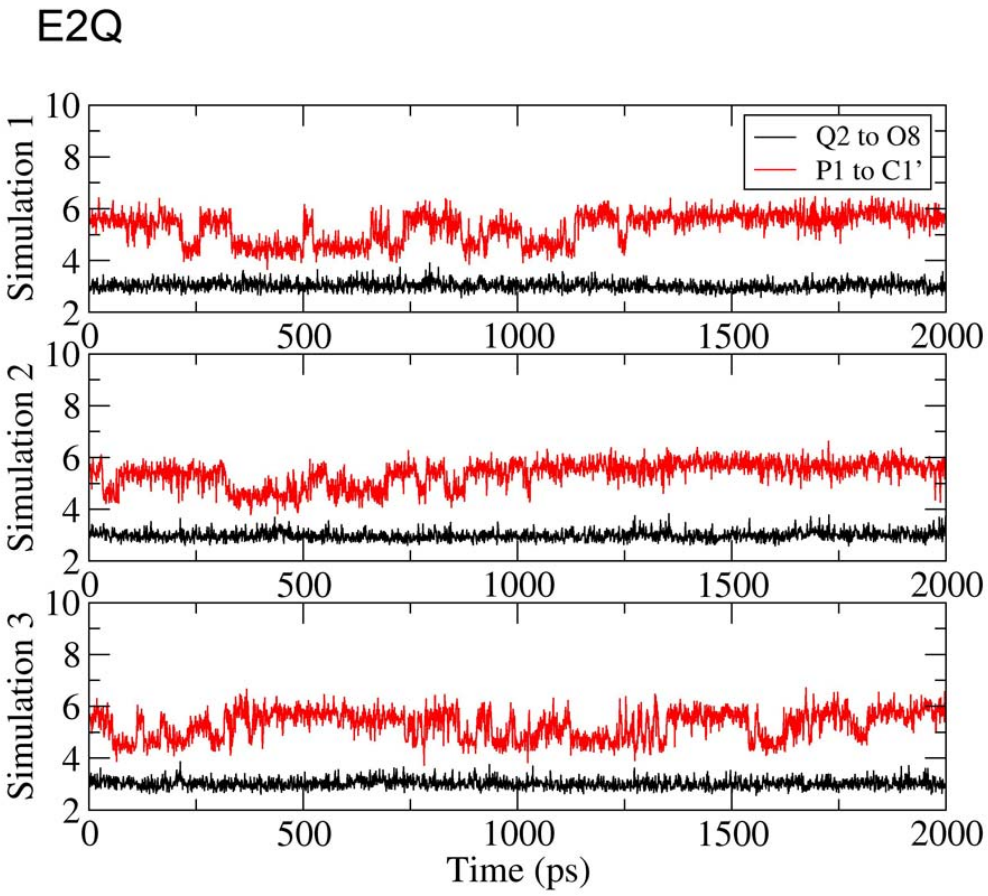

Figure 3-6. The Distance analysis results for E2Q Fpg simulations, in which two distances are shown: the distance between N of catalytic residue Pro1 and C1' of 8OG (labeled as P1 to C1'), and the distance between Nε of Q2 and O8 of 8OG (labeled as Q2 to O8).

The second effect observed is that the average distance between N of Pro1 and C1' of 8OG is shorter in the wild type than in the E2Q mutant. In the proposed mechanism of 8OG excision, C1' of 8OG undergoes nucleophilic attack by the N of Pro1 (*20,*

21). In simulations of wild type Fpg, the average distance is ~4.2 Å. The E2Q mutant samples two minima at 4.5 Å and 5.8 Å, with a preference for the longer distance resulting in an average of ~5.3 Å (Figure 3-5). In the crystal structure, this value is 3.4 Å, much shorter than we observe in the simulation of the same sequence. This apparent inconsistency arises from ~180º rotation of the Pro1 ring in the simulation such that the nitrogen moves farther from 8OG. Interestingly, the rotated ring from the E2Q Fpg simulation still overlaps well with the crystal structure (but with exchange of the positions of N and C) and thus the structure would be expected to remain in good agreement with the electron density (Figure 3-4B). Therefore, the different orientation of the Pro1 ring in our E2Q Fpg simulation and x-ray structure could result from the difficulty of distinguishing between carbon and nitrogen from the electron densities.

Thus, while the control simulations of the E2Q mutant reproduced the crystallographic data with high accuracy, analogous simulations on the wild type sequence resulted in a slight rearrangement of the residues in contact with the lesion. These include loss of a hydrogen bond present with Glu2 and shift of Glu2 away from 8OG, resulting in shortening of the average distance between the Pro1 nucleophile and the 8OG sugar ring. These observations are consistent with the proposed catalytic mechanism.

Previous calculations have suggested that the pKa of glutamate in the active site of T4 Endonuclease V can be shifted to a higher value (*134*). However, Fromme and Verdine have suggested that Glu2 in Fpg is unlikely to be protonated at neutral pH(*121*), due to the observation that glutamates at N-termini of α-helices (such as Glu2 in

Fpg) are less likely to be protonated (*135*).We calculated the pKa value for Glu2 and obtained values of 5.7 and 6.8 depending on the method used for the calculation (see Methods). Since this value is reasonably close to neutral pH, we repeated the simulation using Glu2 with a protonated side chain carboxyl group. The resulting structure was similar to that observed in both the E2Q x-ray structure and the E2Q mutant simulation (Figure 3-7), with a hydrogen bond between O8 of 8OG and Glu2 acidic hydrogen analogous to the hydrogen bond observed between O8 of 8OG and Nε2 of Gln2 in the E2Q mutant. Due to the uncertainty of the ionization state of Glu2 and lack of experimental structure data for the wild type sequence with 8OG bound in the active site, we conclude that it is not possible from these simulations to determine which of these active site hydrogen bonding patterns is adopted under physiological conditions.

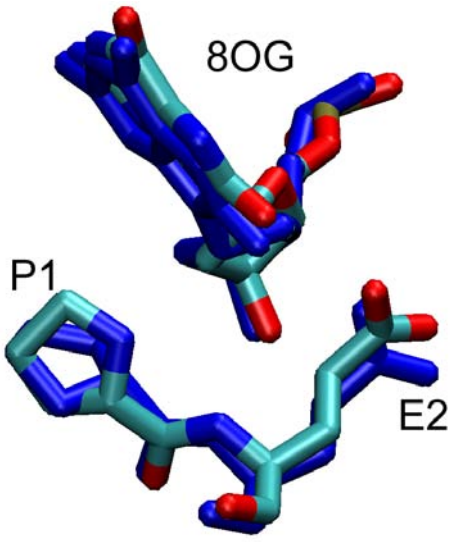

Figure 3-7. Comparison of conformation of the region containing 8OG, Pro1 and Gln2 (Glu2 in wt) between the crystal structure wild type Fpg simulation. In

this simulation the Glu2 was modeled in its protonated state. The crystal structure is shown in dark blue, and the simulation structure is colored by atom.

3.3.4 Structural and energetic analysis of the *anti* and *syn* 8OG binding modes

The glycosidic angle of 8OG in the simulations described above was observed to depend somewhat on the sequence; the glycosidic angle changed from high *syn* (108°) in the E2Q mutant simulation (101° in the x-ray crystal structure) to *syn* (57°) in the wild type sequence. However, a full transition to *anti* was not observed in any of the six simulations. Two scenarios could account for this observation; either the *syn* 8OG conformation is thermodynamically favored, or 8OG also can bind in an *anti* conformation, but the energy barrier between these two conformations is sufficiently large that transitions occur more slowly than the nanosecond time scale of our simulations. Thus we cannot infer from solely this data whether the 8OG's conformation is more favorable in *syn* or *anti*.

To directly compare the stability of the alternate binding modes in the wild type sequence, we constructed a complex containing 8OG in the *anti* conformation by rotation of the glycosidic angle. Simulations of the wild type sequence, for both *syn* and *anti* conformations, utilized two independent methods, umbrella sampling and MM-GBSA, to estimate the relative free energy of these binding modes. These methods differ significantly in their approach and thus provide a measure of the reliability of the conclusions. In addition, umbrella sampling has the advantage of being able to estimate

the barrier for rotation about the 8OG glycosidic bond in the complex, while MM-GBSA can readily provide estimates for the contribution of different interactions to the relative free energies of the two conformations. Thus the methods are independent but highly complementary. We first examine the relative stability of the conformations (since the *anti* form has not been observed crystallographically) and then examine the relative free energies.

3.3.5 Stability of the anti 8OG and syn 8OG binding modes

Simulations were performed for 7 ns for each of the two systems (*anti* and *syn* for wt Fpg). As shown in Figure 3-8, all conformations are stable during the fully unrestrained simulations, with plateau RMSD values of ~1Å for the proteins and DNA fragments. The RMSDs of the loop region (residues 222-231) fluctuate between 1 and 2 Å, with greater flexibility apparent in the *anti* 8OG systems.

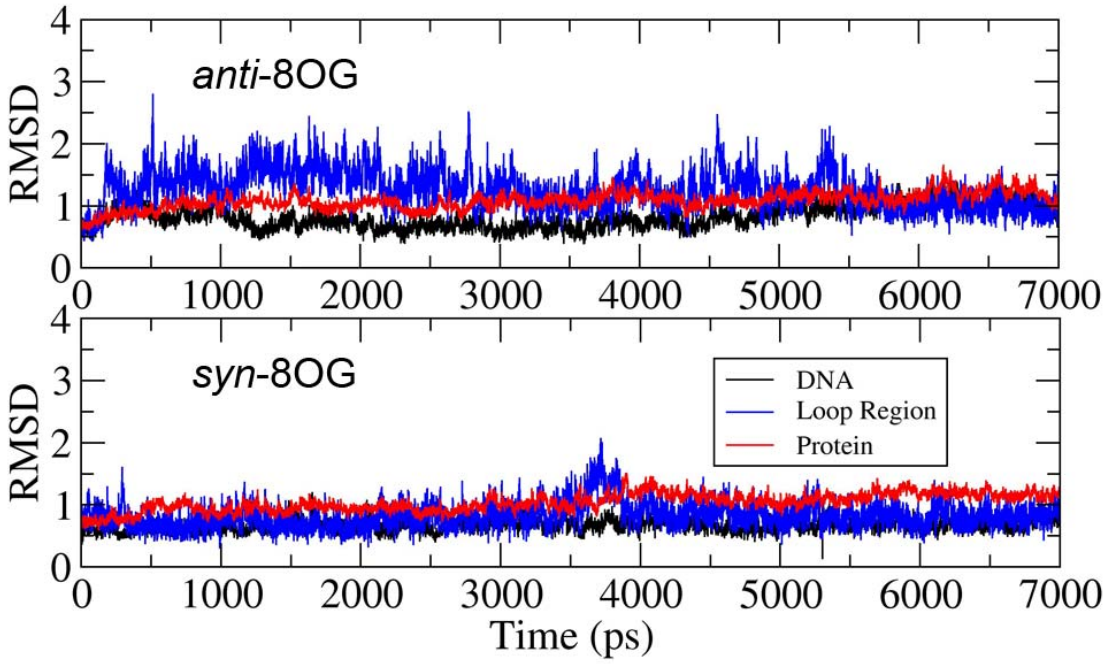

Figure 3-8: RMSD values for the protein, DNA and binding loop during simulations of the WT Fpg/DNA complex with anti (upper) and syn (lower) conformations for 8OG.

3.3.6 Specific interactions between 8OG and Fpg in the two binding modes

Figure 3-9 shows the *anti* and *syn* binding modes of 8OG in the final structures from the simulations. Comparison of the structures reveals surprising similarity in the hydrogen bonding patterns of the two binding modes. Both conformations of 8OG fit in the binding pocket without significantly changing the conformation of the βFα10 loop. In the *syn* conformation (Figure 3-9a), 8OG forms four hydrogen bonds to Fpg: between 8OG N1 and T223 Oγ, between 8OG N7 and the backbone O in S219, between 8OG N2

and E77 Oε and a network of hydrogen bonds between 8OG N6 and the N atoms of residues 221 to 224.

In the *anti* conformation (Figure 3-9b), because of the rotation of the base about the glycosidic bond, N1 and N7 swap their respective partners, with N7 now forming a hydrogen bond with T223's Oγ and N1 hydrogen bonding to the O of S219. N2 acts as hydrogen bond donor to a different Glu residue, E5. N6 maintains its hydrogen bonds with the four residues 221 to 224.

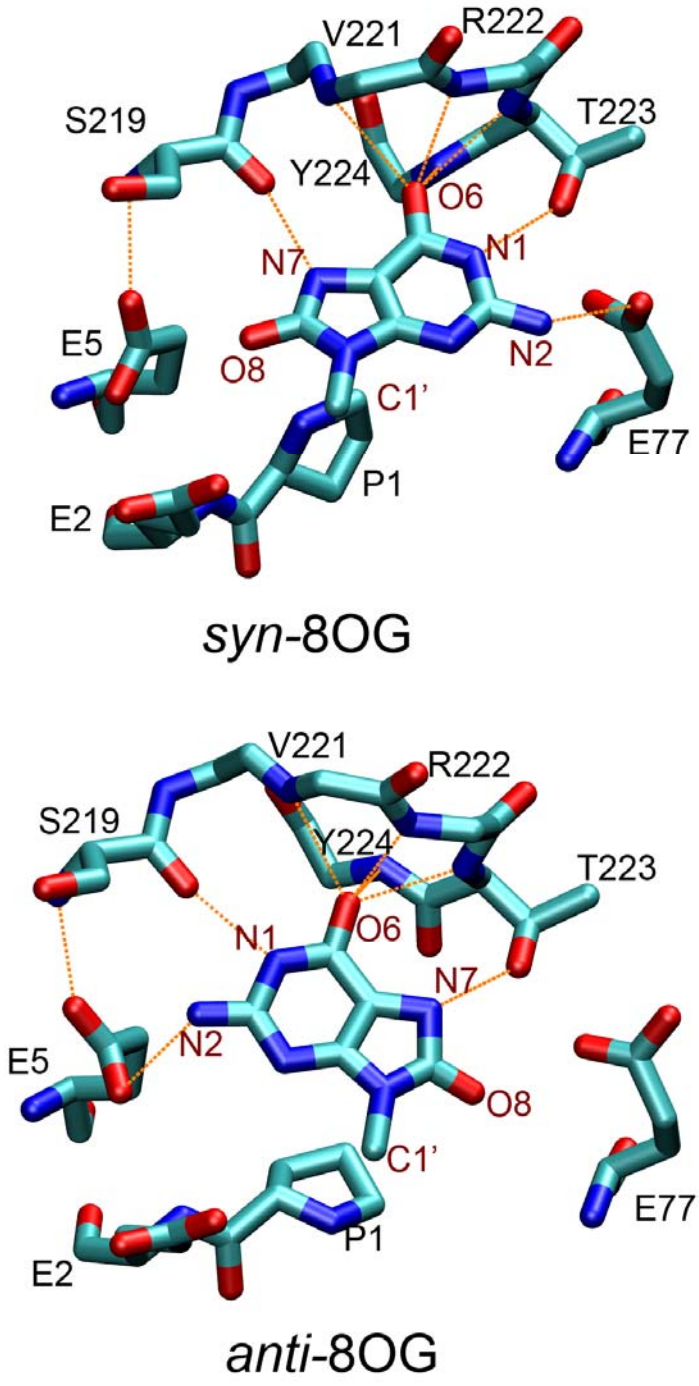

Figure 3-9: 8OG and surrounding residues in the Fpg-DNA complex with (a, left) syn 8OG as observed in the crystal structure and (b, right) anti 8OG built

by rotation around the glycosidic bond of 8OG. Protein residues are labeled in black, and atoms of 8OG are labeled in maroon. Hydrogen bonds are indicated by orange dashed lines. Only the base group of 8OG is shown (the remaining atoms linked to C1' are not shown).

To examine the stability of these interactions during our simulations, we used histogram analysis to compute the distributions of the distances corresponding to these contacts. The results, shown in Figure 3-10, confirm that these hydrogen bond interactions are stable throughout the simulations. In both conformations, 8OG O6 forms a hydrogen bond network with the N atoms of V221, R222, T223 and Y224. For simplicity, the average distance between O6 and these four N atoms is denoted "VRTY". The distributions of the average distances between O6 and the four "VRTY" residues are similar in the two binding modes.

Figure 3-10 also shows that the distribution of the distance between S219 and N7 in *syn* 8OG is similar to that for S219 and N1 in the *anti* 8OG binding mode. This also is the case for the distances involving T223, reflecting the exchange of these hydrogen bonds resulting from rotation about the glycosidic bond (Figure 3-9). 8OG N2 forms a slightly shorter hydrogen bond with E77 in *syn* 8OG than it does with E5 in *anti* 8OG. The longer hydrogen bond distance for 8OG-E5 may arise from the interaction of E5 with S219 in addition to 8OG, while E77 does not form hydrogen bonds with residues other than 8OG and therefore has more freedom to optimize its interaction with 8OG.

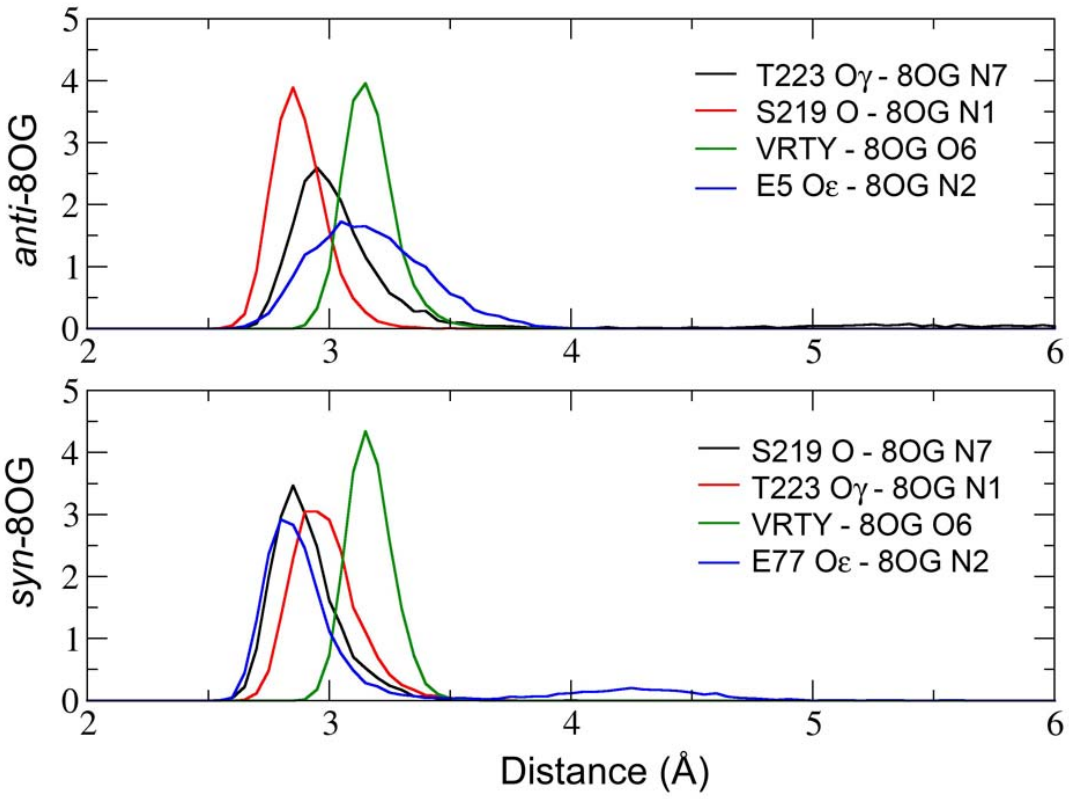

Figure 3-10: Histograms of distances corresponding to hydrogen bonds between Fpg and 8OG. "VRTY" represents the average distance between 8OG O6 and the backbone N atoms in residues 221 through 224.

8OG differs from guanine in hydrogen bonding ability by having an acceptor O8 and conversion of N7 from acceptor to donor. Specific interactions with either group could be employed by Fpg as a final step in allowing the enzyme to discriminate between 8OG and deoxyguanine. In the crystal structure reported by Fromme and Verdine, 8OG N7 in the *syn* conformation forms a specific hydrogen bonding interaction with S219 (*121*). In

contrast, previous simulations of *E. coli* Fpg (*125, 126*) suggested that the O8 of *anti* 8OG interacts with K217 (R222 in *B. st.* Fpg).

Interestingly, the binding pocket in our simulations appears not only able to accommodate both *anti* and *syn* 8OG, but a similar set of specific hydrogen bonds are formed in each. Both conformations have hydrogen bonds that could be used to differentiate 8OG from guanine. In our simulations, the hydrogen bond interaction formed to N7 of *syn* 8OG involves S219, consistent with the crystal structure reported by Fromme and Verdine(*121*). In the *anti* conformation, T223 is the key residue in recognizing the oxidized base, and not R222. Experimental data indicates that 8OG binding affinity is reduced approximately fourfold in the K217T *E. coli* Fpg mutant (*125*). We calculated the distance between R222 and O8 during the simulations, and found no significant interaction. However, since R222 is next to T223, mutation of the basic residue in this position may affect the conformation of T223 and thus the binding affinity of 8OG.

We also examined the effect of the *anti* vs. *syn* change on the distance between P1 nitrogen and the C1' of 8OG. As described above, P1 N acts as a nucleophile attacking C1' of the 8OG in the initial step of the glycosylase reaction (*20, 21*). The distance between these two atoms is significantly different in the two binding modes. In *syn* 8OG, this distance is 4.2 ± 0.3 Å. In *anti* 8OG, this distance is 3.5 ± 0.3 Å. The shorter distance in *anti* 8OG could suggest that this conformation may be more reactive; however other factors are likely to be involved.

3.3.7 Does Fpg preferentially bind *syn* or *anti* 8OG, and what is the role of sequence conservation?

The MD simulations described above demonstrate that 8OG can adopt both *anti* or *syn* conformations in the active site. However, no transitions between these two conformations are observed in these simulations. Similarly, Perlow-Poehnelt *et al.* reported no *anti/syn* transitions in their simulations of *E. coli* Fpg (*126*). Therefore, we cannot directly obtain the relative stabilities of these conformations by comparing their populations. Instead, we used two other approaches, umbrella sampling and MM-GBSA, to calculate the relative stability of each of these conformations in the Fpg binding site.

We also used these MD simulations to investigate the role of the E77 side chain in 8OG binding. Figure 3-9 demonstrates that E5 and E77 appear to have opposite effects for *syn* and *anti* 8OG conformations. E77 stabilizes *syn* 8OG by forming a hydrogen bond with 8OG N2. On the other side of the binding pocket, the E5 side chain has Coulombic repulsion with 8OG O8. In *anti* 8OG, these roles are reversed; E5 forms a hydrogen bond with 8OG N2, and the side chain carbonyl of E77 repels 8OG O8.

Importantly, E5 is strictly conserved while E77 is not. In 85 sequences of Fpg found in the Swiss-Prot database, 48 have serine at the position corresponding to E77, 11 have threonine, and 26 have glutamate. This residue is serine in *E. coli* Fpg. Due to the direct interaction of E77 with 8OG and its different role with *anti* and *syn* 8OG conformations, sequence changes at this position could significantly affect the binding mode. To investigate this possibility, two additional systems were prepared for the E77S Fpg

mutant with *syn* and *anti* 8OG.

Using the umbrella sampling procedure described in Methods, the potential of mean force for rotation of the 8OG glycosidic bond in the Fpg binding pocket was calculated for each of the two sequences (wt and E77S). The results are shown in Figure 3-11, and detailed values including statistical uncertainties are provided in Table 3-1. There are two obvious energy minima in the free energy profiles for both systems. One, located at about 55°, represents the *syn* conformation of 8OG. The other, at about -67°, corresponds to a high *anti* 8OG.

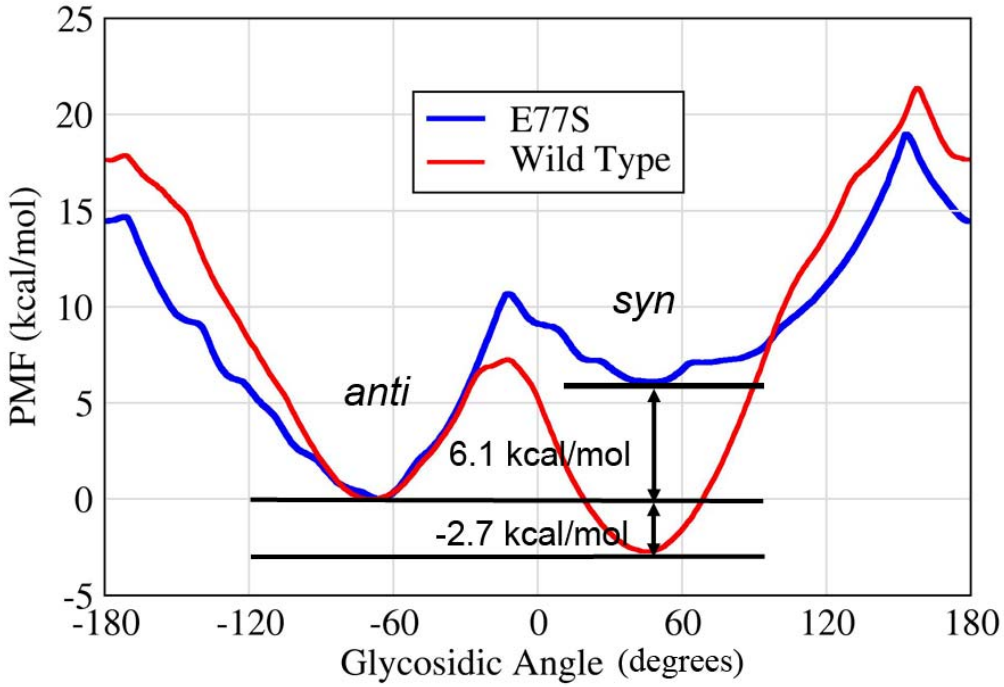

Figure 3-11: The potential of mean force (free energy profile) for rotation around

the 8OG glycosidic bond in the Fpg binding pocket. Data is shown for the B. st. wild type (red curve) and the E77S mutant (blue). To aid the comparison, the free energy of the anti minimum was assigned a value of zero for both data sets.

#Coor	Free	+/-
-42.631579	5.529660	0.000141
-33.157895	7.416580	0.000004
-23.684211	9.566586	0.000002
-14.210526	9.936011	0.000001
-4.736842	9.140676	0.000000
4.736842	6.574496	0.000001
14.210526	3.883372	0.000002
23.684211	1.841038	0.000007
33.157895	0.488213	0.000025
42.631579	0.000000	0.000082
52.105263	0.306151	0.000131
61.578947	1.394883	0.000199
71.052632	3.165009	0.000388
80.526316	5.578786	0.000878
90.000000	8.536848	0.002690
99.473684	11.820639	0.031394

108.947368	14.281074	0.238399
118.421053	16.087819	0.260039
127.894737	18.307050	1.480609
137.368421	19.957345	3.457374
146.842105	21.468789	4.279353
156.315789	23.329062	2308.102730
165.789474	21.904364	3976.671046
175.263158	20.433367	476.737050
184.736842	20.344561	1166.147560
194.210526	19.761588	2009.876684
203.684211	18.744262	2625.359542
213.157895	17.285171	14709.251138
222.631579	14.729963	5434.899463
232.105263	12.590795	2585.929257
241.578947	10.598661	1083.097244
251.052632	8.704132	592.664210
260.526316	6.687245	162.629352
270.000000	4.799523	13.413704
279.473684	3.379707	1.360545
288.947368	2.698482	0.095426
298.421053	2.894129	0.006458
307.894737	4.043388	0.001035
317.368421	5.529660	0.000141

326.842105	7.416580	0.000004
336.315789	9.566586	0.000002
345.789474	9.936011	0.000001
355.263158	9.140676	0.000000
364.736842	6.574496	0.000001
374.210526	3.883372	0.000002
383.684211	1.841038	0.000007
393.157895	0.488213	0.000025
402.631579	0.000000	0.000082

Table 3-1. Output from WHAM analysis on WT-Glu2 umbrella sampling. The first column is the coordinate of the free energy profile (8OG's glycosidic angle). The second column is the free energy value. The third column is the statistical free energy uncertainty.

Although the two energy minima have the same location in wild type and E77S Fpg, the relative free energy of these two minima is highly sensitive to the effect of this mutation. This is consistent with the interactions shown in Figure 3-9, in which E77 makes favorable hydrogen bonding interactions with *syn* 8OG but shows unfavorable Coulombic repulsion with *anti* 8OG. With E77, *syn* 8OG is 2.7 kcal/mol more stable than *anti* 8OG, consistent with observation of the *syn* conformation in the crystal structure of this sequence. In the E77S mutation, however, *anti* 8OG becomes 6.1 kcal/mol more

stable than *syn*. Thus the E77S mutation has an important impact on the relative stability of the two binding modes.

As described above, the ionization state of Glu2 affected the local conformations of the active site. Therefore, we investigated whether the Glu2 protonation would affect the free energy profile for rotation of 8OG by repeating the umbrella sampling calculation with protonated Glu2 in wild type sequence and compared with the profile obtained with deprotonated Glu2. We observe that the ionization state of Glu2 has little effect on the relative free energies of *anti* and *syn* or on the height of the transition state connecting these minima (at glycosidic angles near zero) (Figure 3-12). Larger differences are observed near the highest energy barrier (~150°) where the statistical uncertainties in the data are large (Figure 3-12).

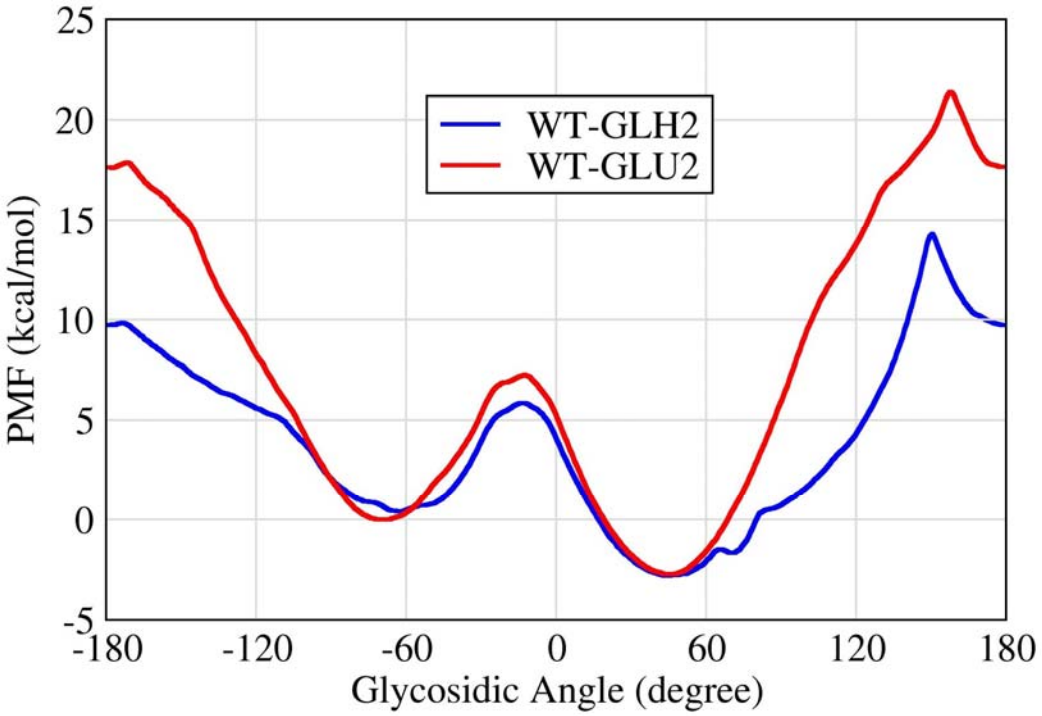

Figure 3-12. The potential of mean force (free energy profile) for rotation around the 8OG glycosidic bond in the Fpg binding pocket. Data is shown for the *B. st.* wild type with Glu2 deprotonated (red curve) and with Glu2 protonated (blue). To aid the comparison, the free energy of the anti minimum was assigned a value of zero for both data sets.

To further confirm these data from umbrella sampling calculations, MM-GBSA was also used to calculate the relative free energies of *anti* and *syn* binding by Fpg using deprotonated Glu2. The advantage of MM-GBSA is that it is possible to obtain approximate contributions to the total free energy from various types of interactions,

including van der Waals, electrostatics, and solvation free energy. Results of this analysis are provided in Table 3-2.

	Wild Type	E77S
ΔΔE(VDW)	2.2±0.1	2.0±0.1
ΔΔE(EEL)	0.4±0.8	-11.5±0.5
ΔΔG(POL)	1.1±0.1	1.0±0.5
ΔΔG(SASA)	-0.1±0.0	0.1±0.0
ΔΔG(total)	3.7±0.6	-8.3±0.1

Table 3-2. The relative anti and syn 8OG binding free energies (kcal/mol), calculated by MM-GBSA, in both B. st. wild type Fpg and the E77S mutant. Positive values indicate that the anti conformation is higher in free energy and therefore less favorable. Total free energies are provided as well as individual components of the free energy. Data are discussed in the text. (VDW: van der Waals, EEL: electrostatic, POL: electrostatic component of solvation free energy, SASA: nonpolar solvation energy using the solvent accessible surface area).

The results of the MM-GBSA calculations are in good agreement with those obtained by umbrella sampling. For the wild type *B. st.* Fpg, *syn* 8OG is 3.7 kcal/mol more stable than *anti* 8OG (2.7 kcal/mol from umbrella sampling). For the E77S mutant,

anti 8OG is 8.3 kcal/mol more stable than *syn* (6.1 kcal/mol from the umbrella sampling calculations).

In addition to providing additional support to the umbrella sampling results, MM-GBSA also supplies estimates of the contribution from each energy term. In wild type *B. st.* Fpg, the total *anti*/*syn* energy difference arises mainly from van der Waals interaction (2.2 out 3.7 kcal/mol). The solvation free energy also favors *syn* 8OG by ~ 1 kcal/mol. The electrostatic energy does not favor either, which is reasonable since there are the number of hydrogen bonds in each conformation are the same(Figures 5 and 6).

The MM-GBSA energy decomposition for the E77S mutant is also shown in Table 3-2. The differences from van der Waals, solvation and solvent accessible area between the two binding modes are roughly the same as were observed with E77. However, the electrostatic energy difference has changed dramatically, with the *anti* conformation stabilized by electrostatic interactions, ~11 kcal/mol more than *syn*. This result is consistent with the discussion above concerning changes in the role of E77 between *anti* and *syn*. The electrostatic energy in E77S favors *anti* 8OG so strongly that it becomes the preferred conformation, despite slightly less favorable van der Waals and solvation free energy.

3.4 Summary and Conclusions

X-ray crystallography continues to be the main source of information for

understanding mechanisms of DNA damage recognition at the molecular level. The need for high quality, diffracting crystals and the requirement of catalytically inactive damaged DNA-protein complexes are frequently addressed by the introduction of single amino acid mutations or, even, by the deletion of a short protein segment. The perturbations that such changes may cause to the structure of the complex can be very difficult to evaluate. In contrast, computational methods are ideally suited to evaluate these putative mutation-derived perturbations by comparing the structures of complexes with the mutated and wild type proteins.

We have focused on the role of the Fpg inactivating E2Q mutation on the specific interactions involved in binding and, perhaps excision of 8OG. Simulations of a complex between the E2Q mutant and DNA provided results consistent with the crystal structure of the same sequence. Simulations of the complex with wild type E2 sequence produced very similar results, with small changes arising from loss of the group on Q2 that hydrogen bonds to 8OG in both the E2Q crystal structure and simulations. The glycosidic angle of 8OG changed from 108° (E2Q) to 57° (wt) and the N-terminal proline moved closer to 8OG, reducing the distance between the P1 nitrogen and the 8OG C1'. In both wild type and E2Q, the *syn* conformation of 8OG and the other key interactions between 8OG and Fpg were maintained as seen in the crystal structure.

We built a model of the wt *B. st.* Fpg-DNA duplex containing *anti* 8OG by using the crystal structure of the complex with *syn* 8OG and rotating the 8OG glycosidic bond. Subsequent unrestrained molecular dynamics simulations resulted in the interesting observation that the binding site (including the βFα10 loop) can readily

accommodate both *anti* and *syn* 8OG without significantly changing its conformation. Additionally, hydrogen bonds seen in the crystal structure and simulations of the *syn* 8OG complex were replaced by a nearly equivalent set for *anti* 8OG. Free energy calculations showed that *syn* 8OG is 2.7 kcal/mol more stable in this binding site than *anti* 8OG.

As suggested by early studies(*134*), the pKa value of Glu2 can have a large shift towards protonated state. Therefore, we also simulated the complex with protonated Glu2. The results indicate that the details of the hydrogen bonding between 8OG and Glu2 depend on the ionization state; however, its effect on the relative free energies of 8OG's two binding modes is not significant (Figure 3-12).

Previous studies also suggested that 8OG could possibly be accommodated by Fpg in either the *anti* or *syn* conformation but that *anti* is favored(*125, 126*), based on simulation of *anti* and *syn* 8OG that were modeled into the *E. coli* Fpg crystal structure with a homology-modeled substrate binding loop. In those simulations, the *anti* conformation was shown to make stable contacts that were consistent with existing mutation data, while the *syn* conformation did not. The authors suggested that the E2Q mutation used for crystallization was one possible reason for the disagreement between their results and the crystal structure. In the present study, we find that the *syn* 8OG conformation that is preferred with wt *B. st.* Fpg arises from favorable interaction with the non-conserved E77 side chain, which also destabilizes the *anti* conformation. This position in *E. coli* Fpg and many other Fpg sequences is occupied by serine, which has a shorter, neutral side chain. We hypothesized that the residue at this position could alter the preferred conformation

of bound 8OG, and tested this postulate through free energy calculations using the *B. st.* E77S mutant. These revealed that *anti* 8OG is indeed more stable than *syn* when E77 is replaced by serine.

While our simulations resolved an apparent discrepancy between two proposed models of binding of 8OG by Fpg, the simulations do not reveal any evolutionary advantage for using different binding conformations by *E. coli* and the thermophilic *B. st.* Moreover, the effect that the *syn* or *anti* 8OG binding site conformation, including hydrogen bonds specific to 8OG, has on the mechanism of lesion extrusion and/or damage recognition remains to be established. Future computations will address these points. It is interesting to note, however, that our simulations demonstrated that the *anti*/*syn* conformational change that makes a large difference in the properties of duplex DNA appears to make relatively little difference in the ability of Fpg to specifically bind the lesion.

Chapter 4 Molecular Mechanics Parameters for the FapydG DNA lesion

4.1 Introduction

Cellular DNA is constantly facing oxidative stress from both exogenous and endogenous resources, which damages DNA by creating oxidative lesions. Failure to repair these lesions can lead to aging related diseases, including cancer(*115, 145*). Among these lesions, 8-oxo-guanine (8OG)(*116*) is one of the most common forms(*11*). Failure to repair the damaged base can cause G:C to A:T transversion(*22*). 2,6-diamino-4-hydroxy-5-formamidopyrimidine (FapydG), another common form of oxidative lesion, shares the same precursor as 8-oxodG (Figure 4-1)(*146*). Both DNA lesions can be excised by DNA glycolases in prokaryotes and eukaryotes(*147*). Because of the increased conformational freedom due to the imidazole ring rupture, it is important to compare the mechanisms of recognition of FapydG and 8OG by the corresponding glycosylase enzymes. However, due to the opened imidazole ring, FapydG tends to anomerize and decompose under conditions required for in vitro DNA synthesis, which will reduce the purity of the synthesized oligonucleotides containing this lesion. This increases the difficulty in studying FapydG. Therefore, unlike the many studies have been done on the recognition of 8OG, relatively little structural information has been obtained about how

FapydG is recognized by DNA repair enzymes.

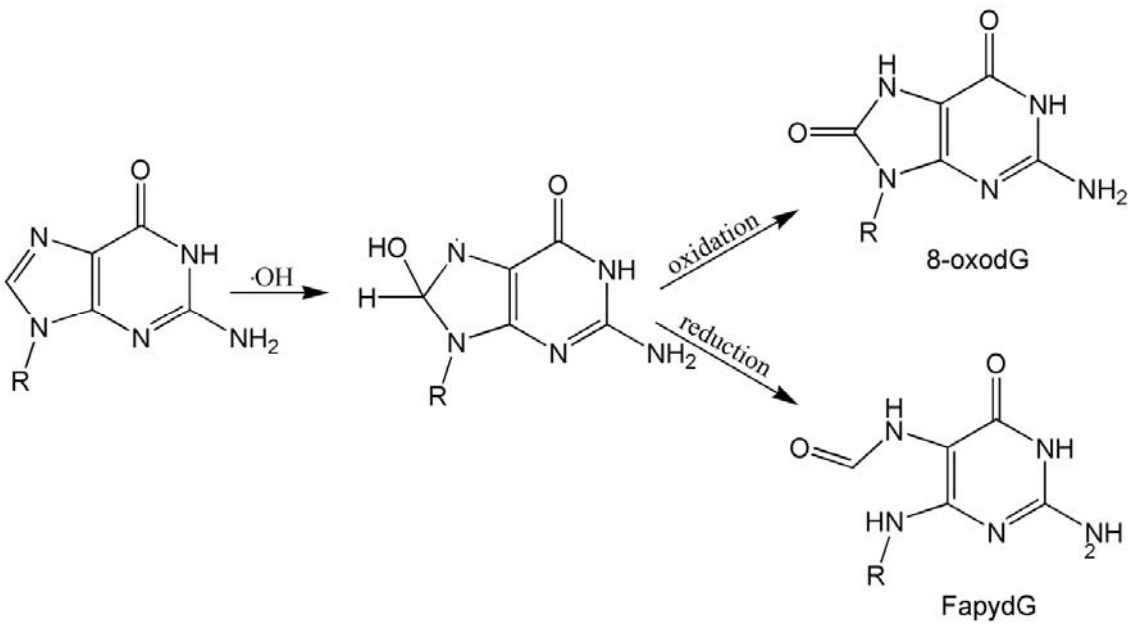

Figure 4-1. The formation of 8-oxodG and FapydG by hydroxyl radicals. Note the loss of the imidazole ring in FapydG-dG that is retained in 8-oxo-dG.

Despite the importance of FapydG, Direct experimental study on three-dimensional structures of FapydG is difficult because it has a high tendency to anomerize and to decompose under conditions required for *in vitro* DNA synthesis. Two types of FapydG analogs in which the ribose is replaced by a pentane ring (*c*FapydG), β-C-FapydG(*148-151*) and carbocyclic FapydG(*152-154*), have been commonly used as a substitute for FapydG since they exhibit increased stability. However, caution must be used when using *c*FapydG data to gain insight into unmodified FapydG behavior.

Molecular mechanics studies can study FapydG directly, and can also provide insight into any differences between *c*FapydG and FapydG, providing a useful framework to interpret *c*FapydG experimental data. This type of simulation depends critically on the accuracy of the molecular mechanics parameters employed. Since FapydG is a non-standard nucleotide, the parameters are not as mature and well validated as those for standard DNA and RNA systems. The major goal of this study is to develop and validate the parameters used in the simulations of FapydG containing systems. The resulting parameters will be used in future simulation studies of the dynamic aspects of FapydG recognition.

Coste *et al.* solved the crystal structure of Fpg from *Lactococcus lactis* (*Ll*Fpg) bound to a *c*FapydG containing DNA(*152*). Although both FapydG and 8OG are excised by Fpg, the *c*FapydG lesion is observed to adopt the *anti* conformation in the Fpg active site while 8OG is observed in the *syn* conformation in the active site of *B. st.* Fpg(*121*). One interesting feature is that *c*FapydG is non-planar (the dihedral angle C4-C5-N7-C8 is -103°). This conformation is stabilized by water-bridged interactions between the O8 of *c*FapydG and side chain hydroxyl of Tyr238, and between the N7 of *c*FapydG and the carbonyl O of Met75. Tyr238 also forms a hydrogen bond with the *c*FapydG phosphate (Figure 4-2).

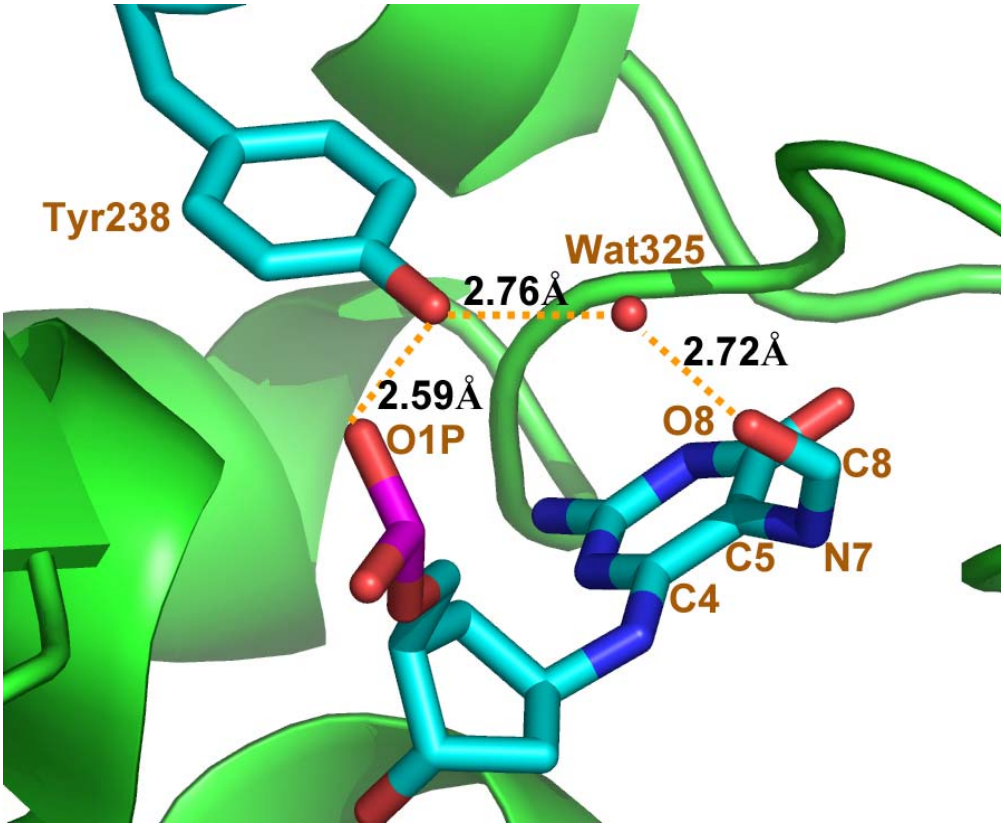

Figure 4-2. The conformation and hydrogen bonds of cFapydG in the X-ray structure (pdb code 1XC8). The hydrogen bonds between FapydG, Tyr238 and Wat325 are shown in orange dashed lines. The heavy atom distances of these hydrogen bonds are also noted.

Computational studies of the recognition of FapydG by the bacterial repair enzyme Fpg have been reported by Perlow-Poehnelt *et al.*(*126*). They carried out 2 ns explicit solvent all-atom MD simulation using Amber ff99 force field(*132*) using two different starting structures; simulations initiated with *anti* FapydG adopted the water bridged interaction between Met73 (Met75 in *Ll* Fpg) and FapydG as seen in the crystal structure.

The resulting simulation model of the interaction of FapydG with Fpg is not consistent with the X-ray structure of *c*FapydG bound to Fpg. One significant difference is the dihedral angle of C4-C5-N7-C8. This dihedral angle in the X-ray structure is -103°. In the simulation with *anti*-trans-FapydG, this value is about 164°(personal commutations? This is OK?) and no interactions involving the O8 of FapydG were observed. It should be pointed out that the sequences of *Ll* Fpg and *E. coli* Fpg are deferent, eg. E76 in *Ll* Fpg is corresponding to S74 in *E. coli* Fpg. The significant effects of this difference has been shown in the other calculation(*78*). It is possible that the difference between the simulation study and x-ray structure is because of the sequence difference. It could also arise from differences between FapydG and *c*FapydG, but it is more likely that the FapydG parameters were not accurate enough to reproduce the effects of a single water bridge. However, this effect could be important for its potential biological relevance.

Although Perlow *et al*. generated new partial atomic charges for the FapydG lesion, the dihedral parameters for bond rotation profiles were adopted from the Amber ff99 force field(*132*) without modification. Due to the opening of the imidazole ring, FapydG has different resonance structures from dG, with loss of aromaticity for the ex-imidazole ring in FapydG. This would be expected to dramatically change the energy profile for rotation about bonds that no longer have significant double bond character, such as C4-C5-N7-C8, suggesting that the dihedral parameters for dG may not appropriate for FapydG. In this study we used ab initio calculations to calculate the energy profile for the rotation of the dihedral angles C4-C5-N7-C8 and C5-N7-C8-O8 (Figure 4-3). The results show that the former has a significantly reduced rotational barrier as compared to what

was defined in ff94/ff99, and the latter is essentially unchanged with significant double bond character. New molecular mechanics parameters were obtained through fitting to these ab initio energy profiles.

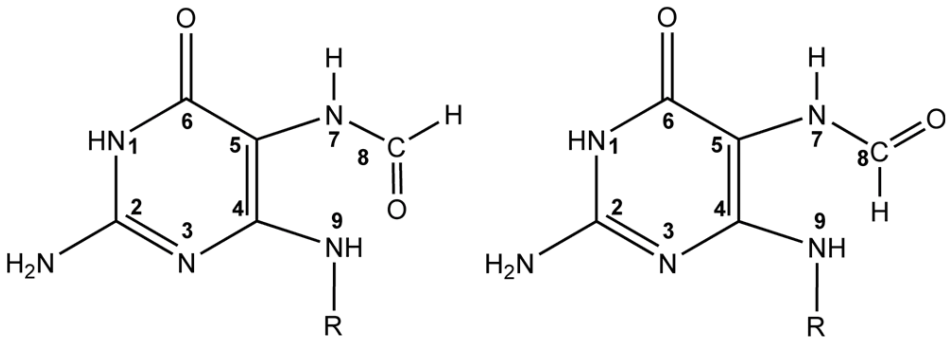

Figure 4-3. The structures of two isomers for FapydG. In cis-FapydG the N7-C8 bond is in the cis configuration, while in in trans-FapydG the N7-C8 bond is trans.

We performed simulations of a DNA duplex containing FapydG, in complex with Fpg from *Bacillus stearothermophilus* (B.st. Fpg). The initial model was obtained using the crystal structure of the complex, with the exception that the experimental structure employed cFapydG (PDB code 1XC812). Using partial charges derived for FapydG along with bond, angle and dihedral parameters obtained from standard ff99, we obtain a planar FapydG that is not in agreement with the cFapydG-containing crystal structure. In

contrast, when the simulations are repeated using the optimized dihedral parameters, the resulting structure is in close agreement with the experimental structural data. This suggests that the crystal structure of the complex containing cFapydG is a reasonable model for the interaction of Fpg with FapydG. Additionally, these simulations provide further evidence that simulations with the new force field parameters will be a useful component of structural studies on FapydG lesions and provide an avenue to explore possible artifacts in experiments arising from the use of cFapydG as a model for FapydG.

4.2 Methods

4.2.1 Calculation of partial atomic charges

The partial atomic charges were calculated following previously published procedures(*155*) in order to be consistent with the ff99 force field. One dimethylphosphate (DMP) and four starting structures of the FapydG nucleoside were used. The initial nucleoside structures included *cis* and *trans* C5-N7 paired with *cis* and *trans* N7-C8 (Figure 4-3). These structures were optimized using Gaussian 03(*156*), with Hartree-Fock calculations and the 6-31G* basis set. A two-step RESP fitting procedure(*155*) was carried out using the program RESP(*157*) in Amber(*51*). Atom types were assigned by analogy. The resulting partial charges and atom type assignments are provided in Table 4-1.

Atom name	Atom type	Partial charge

P	P	1.339981
O1P	O2	-0.845764
O2P	O2	-0.845764
O5'	OS	-0.519671
C5'	CT	0.067080
H5'1	H1	0.052483
H5'2	H1	0.052483
C4'	CT	0.223965
H4'	H1	0.063021
O4'	OS	-0.418929
C1'	CT	0.408208
H1'	H2	0.037353
C2'	CT	-0.014370
H2'1	HC	0.031675
H2'2	HC	0.031675
N9	N	-0.609082
H9	H	0.338636
C4	CA	0.3719540
C5	CF	-0.027042
C6	C	0.5137200
O6	O	-0.579244
N1	NA	-0.481965
H1	H	0.357416

C2	CA	0.635741
N2	N2	-0.840338
H21	H	0.384546
H22	H	0.384546
N7	N	-0.529634
H7	H	0.307277
C8	C	0.6237340
H8	H5	-0.003145
O8	O	-0.569206
N3	NC	-0.513953
C3'	CT	-0.000850
H3'	H1	0.105036
O3'	OS	-0.531572

Table 4-1. Atom types and partial charges for the FapydG residue.

4.2.2 ab initio calculation of rotational energy profiles

The deoxynucleoside FapydG was used to obtain rotational energy profiles. Due to the difficulty of rotating the N7-C8 amide bond, the energy profiles for the rotation of the two rotatable bonds were calculated separately for the two N7-C8 isomers. For the C5-N7 bond, snapshots were generated by performing a relaxed potential energy surface scan in ten degree increments using Gaussian03 at the HF/6-31G* level, repeated for *cis* and

trans conformations of N7-C8 (Figure 4-3), resulting in 72 structures. These structures were optimized using HF/6-31G* with this dihedral angle constrained. For the N7-C8 bond, four conformations with dihedral angle C5-N7-C8-O8 of 0, 90, 180, and 270 degrees were built using Schrödinger maestro and optimized using HF/6-31G*. Energy profiles were calculated using single point energies at the MP2/6-31G* level for the optimized conformations.

4.2.3 Generation of new molecular mechanics dihedral parameters

Molecular mechanics energy profiles were calculated using standard Amber ff99 force field without any dihedral terms for X-C5-N7-X, providing a baseline energy for calculation of the required correction terms. The process was subsequently repeated using ff99 with and without the addition of the new dihedral terms in order to determine the extent to which the combination is able to reproduce the *ab initio* data.

The torsional terms were calculated using the following procedure. 1) The difference between the energy profile from the MP2 calculations described above and the MM energy without explicit dihedral term was calculated. 2) Equation 4.1 was used in a non-linear least-squares fit to obtain parameters that minimize this difference. Since C5 and N7 are expected to employ sp^2 hybrid orbitals, two cosine terms with periodicities of 1 and 2 were employed. In this equation V_1 and V_2 are force constants, ϕ is the dihedral angle and γ_1 and γ_2 are phases.

$$E_{tors} = V_1 \times (1 + \cos(\phi - \gamma_1)) + V_2 \times (1 + \cos(2\phi - \gamma_2)) \qquad 4.1$$

4.2.4 Molecular Dynamics simulations

The starting structure of the Fpg protein from *Bacillus stearothermophilus* (*B.st.* Fpg) with DNA containing FapydG was built based on the X-ray structure 1R2Y(*121*). All water molecules in the crystal structure were retained, including the water bridge formed between FapydG and Tyr238. The Q2 mutation used to inactivate the enzyme was reverted to wild type E2. To simulate the similar environment as in *E. coli* Fpg system, a E76S mutation has also been built(*78*). The structures were minimized for 100 cycles of steepest descent and then solvated in truncated octahedron box with a minimum 6 Å buffer between the box edge and the nearest protein atom. The TIP3P model(*129*) was used to explicitly represent 7356 water molecules. Following previous studies(*125, 126*), the N-terminal proline was modeled as neutral to mimic the stage directly before the reaction. The parameters for neutral N-terminal proline were obtained from Perlow-Poehnelt *et al.* (*126*) All molecular dynamics simulations were carried out with the SANDER module in Amber. The solvated systems were minimized and equilibrated following our previously reported studies on Fpg in complex with DNA containing 8-oxodG(*78*): (i) 50 ps MD simulation with protein and DNA atoms constrained and movement allowed only for water; (ii) five 1000-step cycles of minimization, in which force constants for positional restraints on the protein and DNA atoms were gradually decreased; (iii) four cycles of 5000 steps MD simulation with decreasing restraints on protein and DNA. A final 5000 steps of MD were performed without restraints. The

resulting structures were used in the production runs.

SHAKE(*105*) was used to constrain bonds involving hydrogen atoms. The non-bonded cutoff was 8 Å. The particle mesh Ewald method(*52, 138*) was used to calculate long-range electrostatics. Constant pressure (1 atm) and temperature (300 K) were maintained by the weak coupling algorithm(*139*).

4.3 Result and discussion

4.3.1 Simulation results using the default Amber ff99 parameters

*c*FapydG adopts a non-planar, *cis* conformation in the X-ray structure (Figure 4-4)(*152*). The dihedral angle of C4-C5-N7-C8 is -103°, and the O8 of *c*FapydG interacts with Tyr238 through a water molecule. This interaction may be important for the lesion recognition since O8 is absent in undamaged guanine, but is present for both the FapydG and 8OG lesions that are excised by Fpg. In contrast, FapydG structures observed in previously reported MD simulations on *E. coli* Fpg complex were nearly planar (164° for the *anti*-*trans*-FapydG)(*126*). Our simulations using standard Amber ff99 dihedral parameters that were initiated in the non-planar crystal conformation also spontaneously adopted a planar FapydG structure (Figure 4-4), with a C4-C5-N7-C8 dihedral angle of -5°±14° (uncertainty denotes the standard deviation) (Figure 4-7).

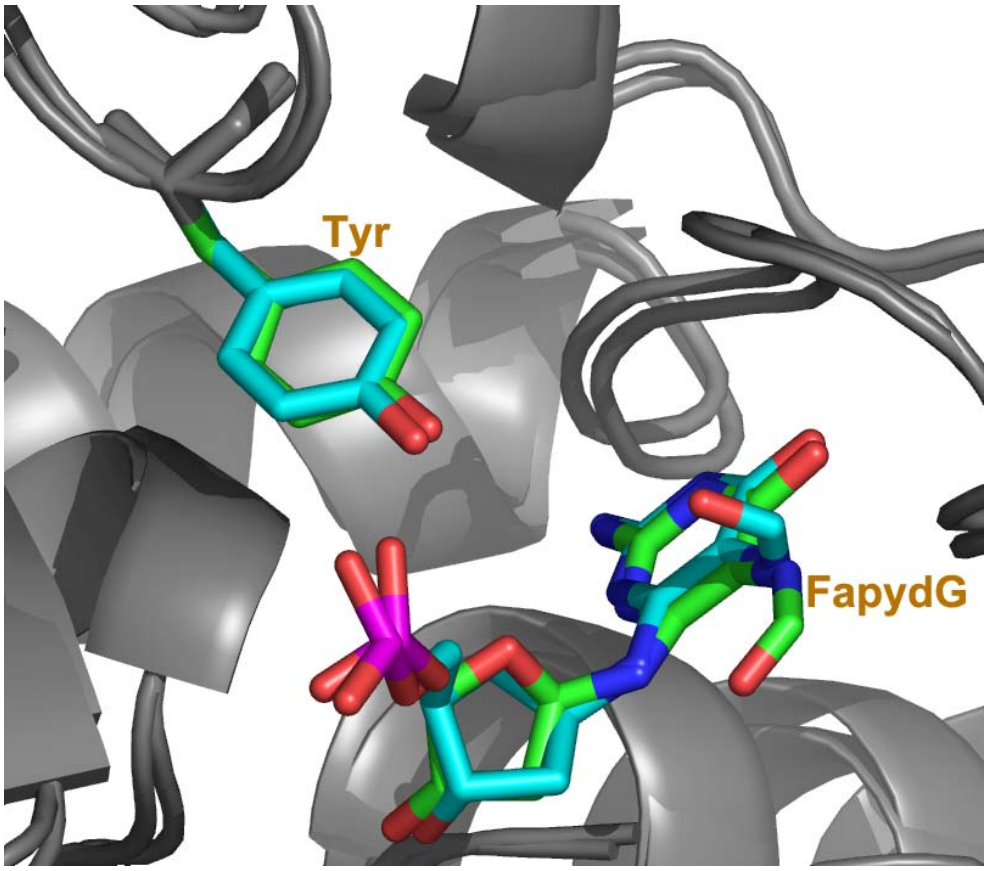

Figure 4-4. The conformation of *L.l.* Fpg bound to cFapydG containing DNA in X-ray structure 1XC8 (carbon atoms shown in cyan) and the simulated conformation of *B. st.* Fpg bound to FapydG containing DNA using standard ff99-based parameters (carbon atoms shown in green). Only the heavy atoms of FapydG and Tyr238 are shown. Although the FapydG simulation was initiated with the crystal structure, the non-planar conformation of cFapydG was not retained when using these parameters.

4.3.2 The energy profiles for C5-N7 bond rotation

The major reason that *c*FapydG is non-planar in X-ray structure is rotation about C4-C5-N7-C8 with a dihedral angle of -103°. In FapydG's closed imidazole-ring analogs, such as guanine and 8oxo-guanine, the C5-N7 bond has partial double bond character due to the aromatic purine ring. The open form of imidazole ring in FapydG can reduce the partial double bond character of the C5-N7 bond, lowering the energy penalty for non-planar conformations to which a water hydrogen bond may be enough to distort the conformation. To test this possibility, we generated the ab initio energy profiles for C5-N7 bond rotation (see Methods). The results are shown in Figure 4-5 for the energy profiles obtained at the MP2 level as well as the profile obtained using standard ff99-based parameters(*125*). The energy profiles are in poor agreement, and the difference between them varies depending on whether the neighboring N7-C8 amide is in the *syn* or *anti* conformation. In particular, ff99 MM energies for non-planar conformations are substantially larger than observed in the MP2 data, suggesting that the dihedral parameters are the source of disagreement between structures obtained from simulation and crystallography.

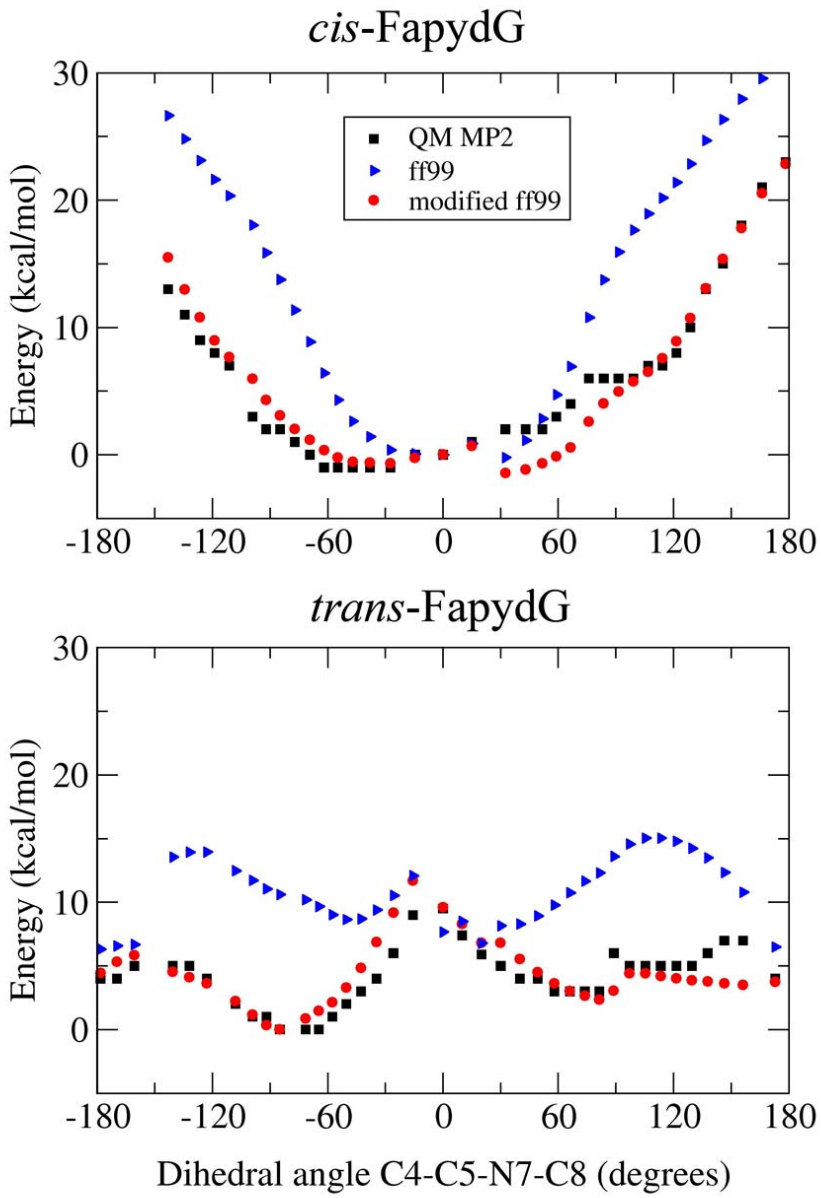

Figure 4-5. The total energies of the molecule with different C4-C5-N7-C8 dihedral angles. The energies were obtained using ab initio calculations (black square), molecular mechanics with torsion parameters obtained from ff99 (blue triangles) and molecular mechanics with ff99 and our modified torsion energy terms (red circle). The new profiles are in much better agreement with the ab

initio data than those obtained using ff99.

To generate new torsional parameters for the C5-N7 bond, we calculated the difference between the energies from the MP2 profile and those obtained from a calculation using ff99 without any dihedral terms for C5-N7 (see Methods). New parameters were obtained through fitting to this difference function, with the resulting parameters shown in Table 4-2 and a comparison of the energy profiles for the original and new parameters shown in Figure 4-6. Consistent with our expectation of reduced double bond character upon ring opening, the rotational energy barrier using the new parameters is about 7 kcal/mol, much lower than the 19 kcal/mol obtained using ff99-based parameters. In addition, the new parameters adjust the relative energies of the minima at 0° and 180° by 2.7 kcal/mol. In ff99, these minima had the same energy values.

	V_1 (kcal/mol)	n_1	γ_1 (degrees)	V_2 (kcal/mol)	n_2	γ_2 (degrees)
ff99	9.6	2	180	-	-	-
New values	3.52	2	180	1.36	1	0

Table 4-2. New and ff99-based torsion parameters for rotation about the C5-N7 bond, using the function provided in Equation 4.1. The dihedral angle is represented by the atoms C4-C5-N7-C8.

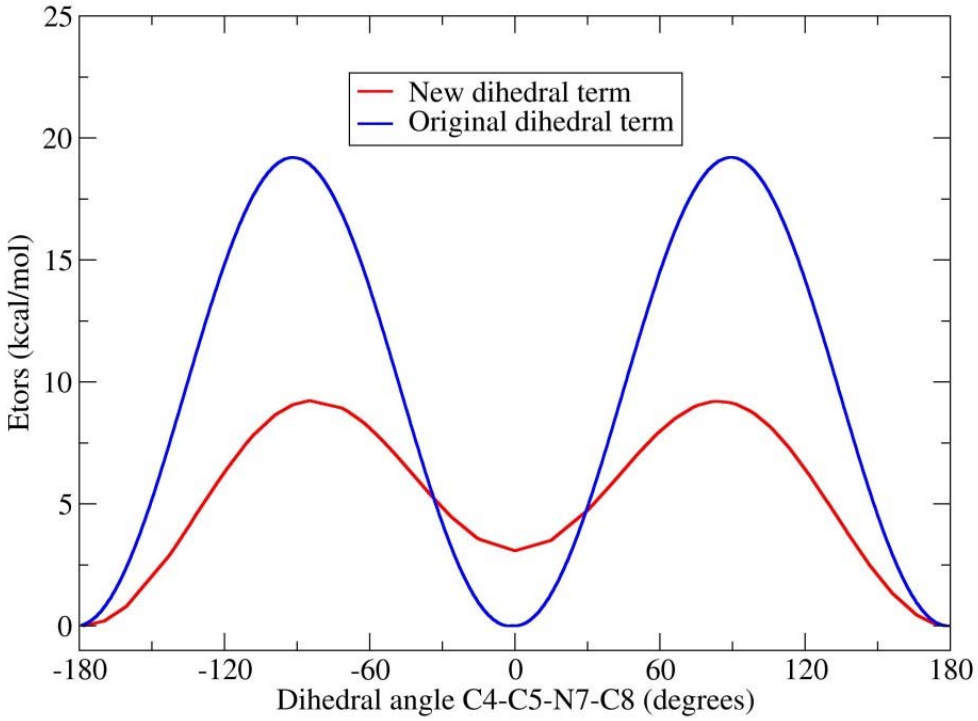

Figure 4-6. Dihedral angle energies for rotation about the C5-N7 bond. Only the dihedral energy (equation 4.1) is shown. The blue line represents standard ff99, and the red line represents the new parameters obtained through fitting to the QM energies.

4.3.3 The energy profiles for N7-C8 bond rotation

Since the barrier for rotation about the C5-N7 bond required modification, we also calculated the energy profile for rotation about the neighboring N7-C8 bond using a

similar procedure as described for C5-N7. The results are shown in Table 4-3. The resulting parameters are provided in Table 4-4, along with those from standard ff99.

Dihedral angle (degrees)	0	90	180	270
Energy: MP2 (kcal/mol)	-2	19	0	29
Energy: MM (ff99) (kcal/mol)	-6.2	14.3	0	28.3
Energy: MM (new) (kcal/mol)	-3.2	17.3	0.0	29.8

Table 4-3. Energies of four FapydG conformations with different C5-N7-C8-O8 dihedral angle and cis C4-C5-N7-C8.

Term	V_1 (kcal/mol)	n_1	γ_1 (degree)	V_2 (kcal/mol)	n_2	γ_2 (degrees)
Value in ff99	10	2	180	2	1	0
New values	10	2	180	-	-	-

Table 4-4. Parameters for the dihedral angle C5-N7-C8-O8 using equation 4.1. The values obtained using ff99 and the new values from the fitting of MP2 single point energies are listed.

In this case the main energy barrier for the rotation around N7-C8 is the same in the new and original parameter sets (20 kcal/mol). A minor difference is that in the new parameter set the *cis* conformation of this dihedral angle is 4 kcal/mol less stable than *trans* conformation. However, this difference may not be noticeable during normal simulations because of the large height of the barrier separating these minima (20 kcal/mol). In our simulations of DNA in solution and in complex no transition between these conformations has been observed.

4.3.4 Results using the new parameters

The molecular mechanic energy of FapydG was calculated using standard Amber ff99 force field with these new torsional energy terms for rotation about C5-N7 and N7-C8. As shown in Figure 4-5, the resulting energy profiles are in much better agreement with the MP2 data, and the profiles for both *syn* and *anti* N7-C8 are reproduced with a single set of parameters.

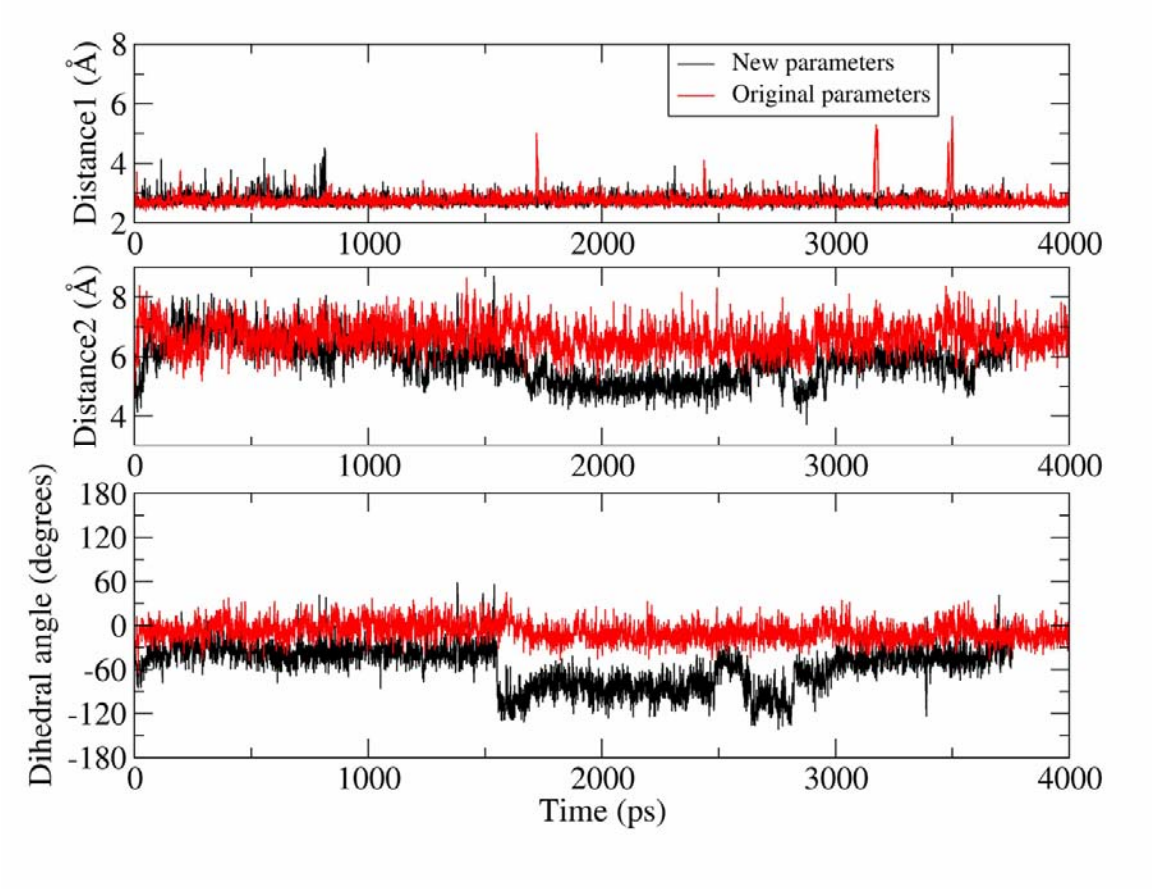

Figure 4-7. The dihedral angles and distances in the B. st. Fpg/FapydG simulation with new and original FF99 parameters. In these figures, distance 1 is the distance of O1P of FapydG to OH of Tyr238, distance 2 is the distance of O8 of FapydG to OH of Tyr238, and dihedral angle is the dihedral angle of C4-C5-N7-C8 of FapydG. In the X-ray structure 1XC8 these values are: the distance1 is 2.59 Å, and the distance2 is 5.08 Å, and the dihedral angle of C4-C5-N7-C8 is -103.74°.

Similar to the simulations described above using standard ff99, simulations of the *B.*

St. Fpg/DNA complex containing FapydG were performed in explicit solvent using the new dihedral parameters. The starting structure for FapydG was obtained from the X-ray structure of *c*FapydG, with a non-planar, cis FapydG conformation. We analyzed the planarity of the FapydG as well as the interaction between FapydG and Tyr241. In Figure 4-7a, we show that a stable hydrogen bond is formed between the OH of Tyr241 and the phosphate O1P atom of FapydG, with an average distance of 2.7±0.2 Å. The distance between FapydG O8 and Tyr241 is much longer (Figure 4-7a); visual analysis revealed that both FapydG and Tyr241 hydrogen bond to a bridging water molecule, although the water exchanges with bulk solvent during the simulation (Figure 4-8). The water bridge between O8 of FapydG and OH of Tyr was lost at the beginning of the simulation but re-formed after ~1500 ps. After that point the distance between O8 of FapydG and OH of Tyr241 was stable at ~ 6Å with fluctuations when the water again exchanged with the bulk at ~ 3000 ps. Figure 4-8 shows the overlap of the two snapshots from different time points in the MD simulation with the X-ray structure of the Fpg/*c*FapydG complex. Not only do FapydG and Tyr214 from our simulation neatly align with the X-ray conformation, but the relatively more mobile water molecules involved in the water bridge also appear at about the same position as the one in the X-ray structure, even though exchanges occurred.

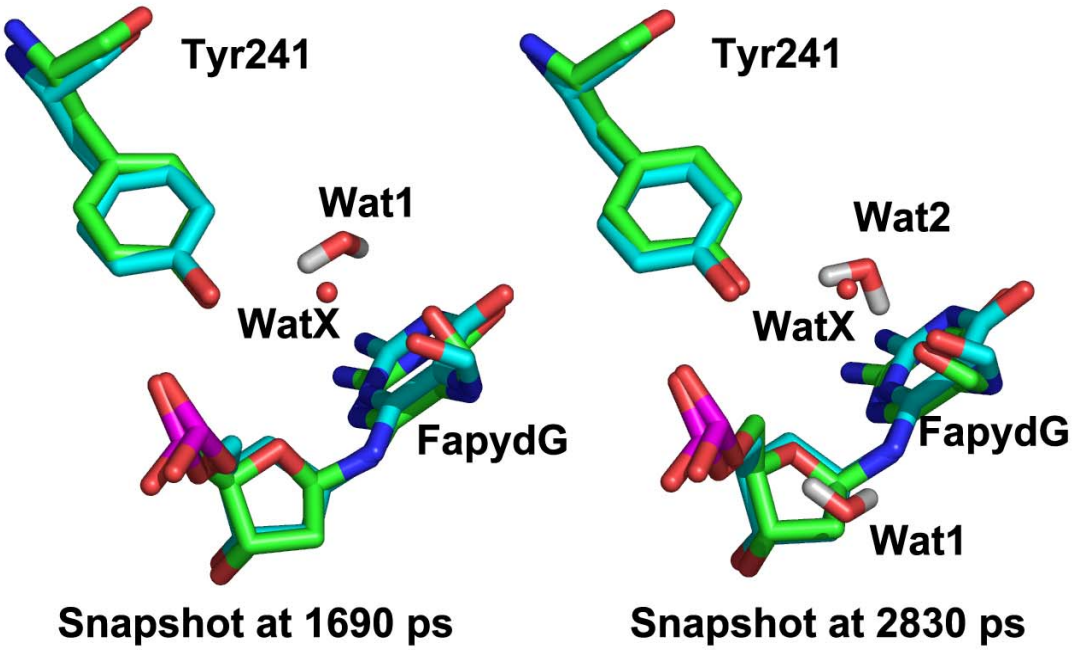

Figure 4-8. Two snapshots from B. st. Fpg/FapydG simulation (carbons in green) overlapped on the X-ray structure of cFapydG (carbons in cyan). Two MD snapshots are shown (1690 ps and 2830 ps), in which two different water molecules (labeled Wat1 and Wat2) play the role in bridging the interaction between FapydG and Tyr241. The FapydG conformation and position of the bridging water molecules are in good agreement with the crystal structure (with water labeled WatX).

In the simulation using the standard ff99 dihedral parameters, FapydG strongly preferred a planar conformation (the dihedral angle C4-C5-N7-C8 is ~ 0°, Figure 4-6), precluding the possibility of formation of this water-bridged interaction.

Figure 4-7C shows the dihedral angle of C4-C5-N7-C8 in FapydG during the simulation. In the simulation using the new parameters, a correlation is apparent between the formation of the water bridge and the value of this dihedral angle. When the water bridge was not stable (0 ~ 1500 ps), the dihedral angle of C4-C5-N7-C8 was closer to planar (-30 °). After formation of the water bridge, (1500 ~ 2900 ps), this dihedral angle was highly non-planar (~-100°), in good agreement with the crystal structure (-103°). This correlation suggests that the formation of the water bridge stabilizes the non-planar conformation, since the energy minimum for the isolated *cis* FapydG occurs in the planar conformation (Figure 4-5). Hence the energy penalty of the non-planar conformation is balanced by the formation of the water bridged interaction. In the simulations using the previous parameter set, the rotation energy barrier, which was overestimated by 10 kcal/mol using ff99 (Figure 4-6), cannot be overcome by the water bridge.

In our previous calculation, we have found that E77 in *B. st.* Fpg has significant effects on the binding mode of 8-oxo-guanine(*78*). Therefore, we also extended our simulation using the new parameter set with E77S *B. st.* Fpg. The results are shown in Figure 4-9. After a short time of equilibration, the dihedral angle of C4-C5-N7-C8 was maintained at about 20° for about 1 ns. Then it switched to -40° and was kept at that value for the rest of the simulation. In this simulation, there was no water bridge formation observed. The water bridge and severe non-planar FapydG conformation which seem important in the substrate binding in *Ll* Fpg (x-ray structure 1XC8) and *B. st.* Fpg (in our simulations, Figure 4-7) could be species dependent.

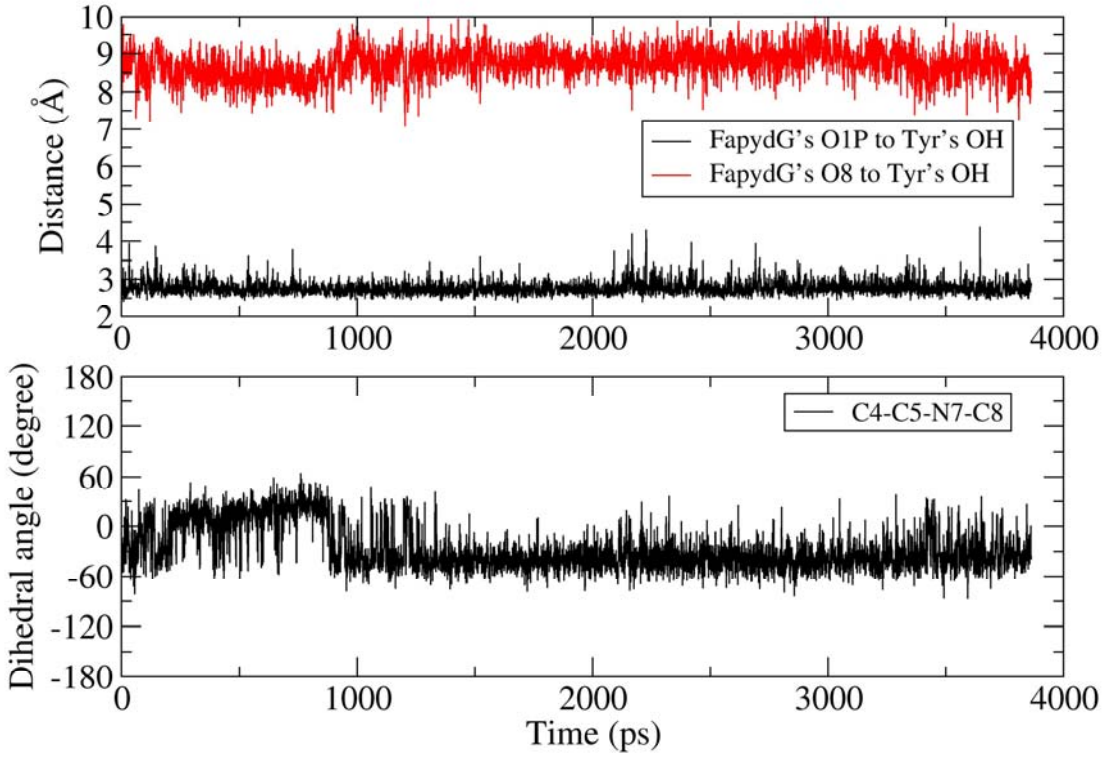

Figure 4-9. Dihedral angles and distances in the E76S *B. st*. Fpg/FapydG simulation using the new dihedral parameters. In the X-ray structure the distance of O1P of FapydG to OH of Tyr238 is 2.59 Å, O8 of FapydG to OH of Tyr238 OH is 5.08 Å, and the dihedral angle of C4-C5-N7-C8 is -103.74°.

4.4 Conclusion

We performed quantum mechanical and molecular mechanical calculations which confirmed that different molecular mechanics dihedral parameters are required for the opened ring of FapydG. This increases the flexibility of the dihedral angle C4-C5-N7-C8

through a reduction in the effective dihedral barrier by about 10 kcal/mol. This effect was not included in previously reported simulations, providing an explanation for the planar conformation of FapydG in those simulations, in disagreement with crystallographic data for cFapydG. The new parameters more faithfully reproduce both the positions of energy minima and the also the height of the barriers that are obtained through *ab initio* calculations. When used in simulations of a FapydG-containing DNA duplex bound to the DNA glycosylase Fpg, the new parameters reproduced the non-planar conformation for cFapydG observed in a crystal structure of the Fpg/DNA complex. This conformation was observed to be stabilized through a water-bridged interaction with Tyr241 in Fpg, which may be important for the specific recognition of the FapydG lesion that is excised by Fpg. Simulations using standard ff99 parameters for FapydG were unable to reproduce this conformation. The new parameters are expected to be a crucial component of future studies of the important FapydG lesion, which is particularly difficult to study using experimental synthesis techniques

Chapter 5 DNA sliding and flipping in the Fpg/DNA complex

5.1 Introduction

Bacterial genomes are composed of over a million nucleotides, and the human genome contains over a billion nucleotides. How DNA glycosylase proteins find DNA lesions in the great excess of normal DNA nucleotides is still one great mystery. It is generally believed that the proteins binds to the DNA double strand randomly and translocates to the lesion site. The exact procedure of the translocation is still under study. There are three proposed searching mechanisms: sliding, hopping, and intersegment transfer(*158*). In the "sliding" mechanism, a protein diffuses along the DNA double strands without dissociation. In this mechanism the searching space is one dimensional, instead of three dimensional. This makes the searching over a short distance very efficient. In the hopping mechanism, the protein completely dissociates from the DNA, diffuses a short distance in the solution, and rebinds to another location on the DNA. By doing this, the protein can "jump" to the second location which could be far away along the DNA sequence, but close in the three-dimensional space. In the third mechanism, "intersegment transfer", the protein has two distinct DNA binding sites, which can bind to two locations of DNA at the same time. By releasing one binding and reforming a new

binding at a different location, the protein swings among DNA segments without ever dissociating from DNA. This type of mechanism needs the protein to have two DNA binding sites. A recent study done by Gower *et al.* has shown that translocation is a mixed mechanism(*159*). Under *in vivo* salt conditions, the transfer shorter than 30 bp is mainly sliding, and the one over 30 bp will include at least one step of dissociation/rebinding step. Due to the limitation of computer power, we focused on the study of sliding over 1 to 2 bp.

Our research is guided by the novel hypothesis that lesion "recognition" is a complex, multi-step procedure, which is activated when DNA glycosylases bind to damaged DNA(*30*). This procedure starts as a nonspecific binding of DNA glycosylases to DNA duplex, and rapid sliding along the duplex till they reach the lesion site or hop to the other segment of DNA. Lesioned nucleotides flip from the intrahelical position and enter the active site. The formation of the specific interactions between the lesioned DNA and the active site residues are only the last step of the recognition procedure. Each step shown in Figure 5-1 could act as a threshold which only allows or favors the lesioned nucleotides to be processed by the enzyme.

Several recently published x-ray structures have resolved the conformation of the system at different stages. X-ray structure 1K82(*25*), 1L1Z(*27*) exhibit the structure of Fpg covalently bond to the DNA after the base group excision. X-ray structure 1R2Y(*121*), 2F5Q and 2F5S(*160*) show the structure of Fpg/DNA complex with 8OG binding inside the active site. X-ray structure 2F5N, 2F5O, and 2F5P(*160*) show the conformation before the base flipping. Two x-ray structures of hOGG1, the functional

homolog of Fpg in human, present the intermediate states in the base flipping pathway(*161, 162*).

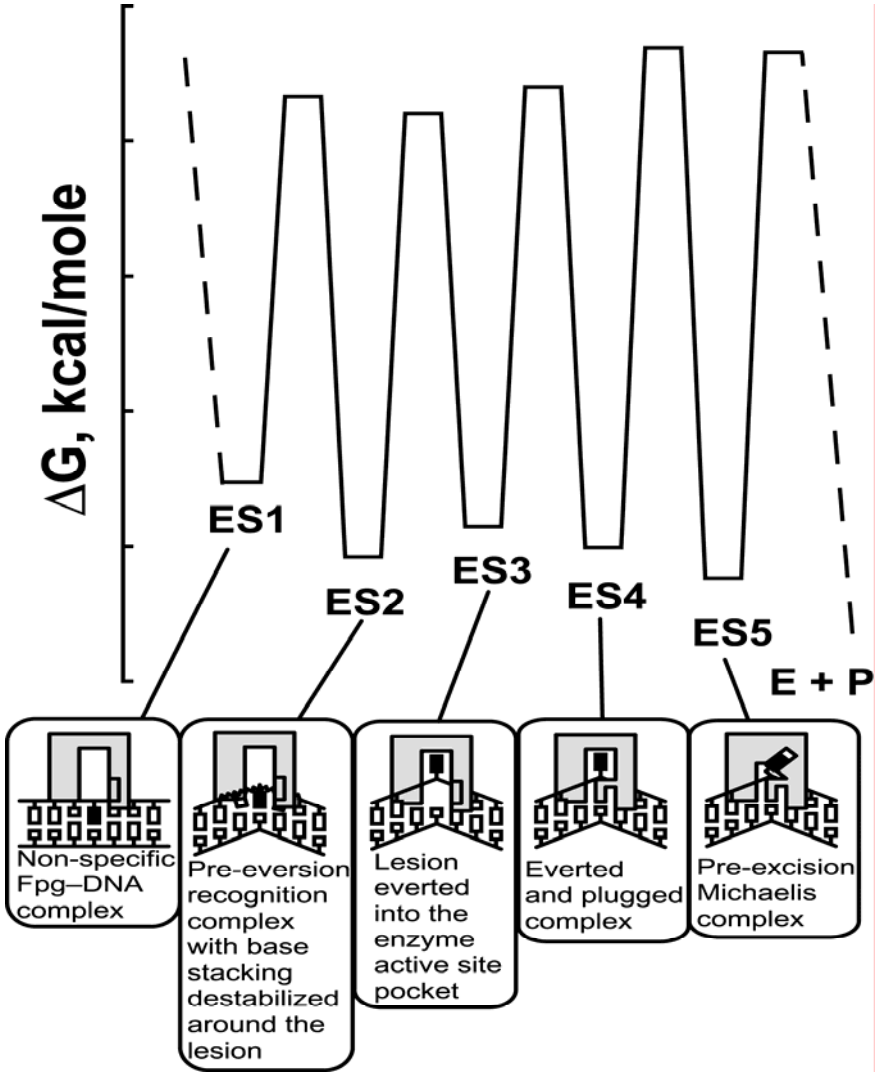

Figure 5-1. Proposed reaction coordinate for Fpg recognizing lesioned nucleotide (*30*). ES1, encounter complex; ES2, recognition complex; ES3, everted complex; ES4, everted and plugged complex; ES5, Michaelis complex.

Ticks at the ΔG axis correspond to 5 kcal/mole intervals.

Our goal is to characterize structurally, energetically and kinetically the discrete steps involved in sliding, lesion binding, base eversion, and isomerization of the enzyme globule. These steps, summarized in Figure 5-1, include ES1, a nonspecific complex between the glycosylase and the lesion or lesion site; ES2, an initial recognition complex with base pairing destabilized around the lesion; ES3, a complex with the damaged base everted into the enzyme's active-site pocket; ES4, a complex with several amino acids inserted into the helix to fill the void created by the base eversion; and ES5, the complex isomerized to form the pre-excision, Michaelis complex (*30*).

5.2 Methods

5.2.1 System preparation

All initial structures were built using the Leap module of Amber (version 9) (*128*), based on the crystal structure of the *B. st.* Fpg crosslinked with DNA (2F5O.pdb)(*160*). The sequence of the DNA duplex was 5'-AGGTAGACCTGGACGC-3', 5'-TGCGTCCG*GATCTACC-3' (where G* is at the interrogating site). All water molecules in the crystal structure were retained. The DNA and protein mutants were generated by manual editing of the pdb file, with the new side chain built using Leap. These structures were minimized for 100 cycles of steepest descent and then solvated in truncated octahedron boxes with a minimum 6 Å buffer between the box edge and the nearest protein or DNA atoms. The TIP3P model(*129*) was used to

explicitly represent water molecules. Following previous studies(*78, 125, 126*), the N-terminal proline was modeled as neutral to mimic the stage directly before the reaction. The parameters for neutral N-terminal proline were obtained from Perlow-Poehnelt *et al.* (*126*). Force field parameters for 8OG were obtained from Miller *et al*.(*130*). Zinc was modeled using the ion zinc finger model (*163*). The remaining protein and nucleic acid parameters employed Amber ff99 (*100, 132*), with modified protein backbone parameters to reduce the alpha-helical bias of those force fields(*133*).

5.2.2 Molecular dynamics simulations

Eight long MD simulations were performed. They were labeled as VIC3 (the structure built based on x-ray structure 2F5O), VOIC3 (the same as VIC3 except that the guanine in the interrogating site was changed to 8OG), F114AGC (the same as VIC3 except Phe114 was mutated to alanine), and F114AOGC (the same as F114AGC except that the guanine in the interrogating site was changed to 8OG). For each system, two simulations were carried out from the same starting structures, named as A and B.

All molecular dynamics simulations were carried out with the SANDER module in Amber. Solvated systems were minimized and equilibrated in three steps: (i) 50 ps MD simulation (*128*) with protein and DNA atoms constrained and movement allowed only for water; (ii) five 1000-step cycles of minimization, in which the positional restraints on the protein and DNA were gradually decreased; (iii) Four cycles of 5000 steps MD simulation with decreasing restraints on protein and DNA. A final 5000 steps of MD were performed without restraints. The resulting structures were used in the production

runs.

SHAKE(*105*) was used to constrain bonds involving hydrogen atoms. The non-bonded cutoff was 8 Å. The particle mesh Ewald method(*52, 138*) was used to calculate long-range electrostatics. To mimic the optimum living temperature for bacteria *B. st.*, All these simulations were performed at 330 K. Constant temperature was maintained by the weak coupling algorithm (*139*).

5.2.3 Anisotropic network model analysis

Anisotropic network model (ANM) is a simple normal mode analysis tool for analyzing vibrational motions in proteins(*164, 165*). It represents the macromolecule as a network, in which each node is the Cα atom of a residue and the overall potential is simply the sum of harmonic potentials between interacting nodes. By doing so it allows prediction of anisotropic motions. This method has been successfully applied for explaining the relation between function and dynamics for several proteins (*166-168*). To aid users, the developers have set up a web interface (http://ignmtest.ccbb.pitt.edu/cgi-bin/anm/anm1.cgi). On the webpage the user need to submit a pdb file or the pdb id from PDB database. The server behind the web interface will excute the calculation and post the result on the webpase for downloading.

5.2.4 Targeted MD simulations

Targeted MD simulation was used to simulate the process of 8OG's base eversion. In

this simulation the structure of the intra-helical 8OG:C base pair was built based on the cross-linked x-ray structure (pdb id: 2F5O), and it was used as the starting structure. The extra-helical 8OG structure was built based on x-ray structure with flipped 8OG base group (pdb id: 1R2Y), and it was used as the targeting structure. Targeting force (force constant was 5 kcal/mol×$Å^2$) was used to force the system to evolve from the starting structure to the targeting structure. The force was applied on all heavy atoms in the base group of 8OG. The targeting RMSD value was changed gradually from the original difference between starting structure and targetting structure to 0 over 500 ps simulation. The input file for this targeted MD simulation is shown in Appendix F – Targeted MD simulation. The protein sequence in both x-ray structures is from bacterium *Bacillus stearothermophilus.* To mimic the optimum temperature of *Bacillus stearothermophilus* bacterium, this simulation was done on 330 K.

We also used targeted MD simulation to mimic the DNA sliding on the Fpg/DNA interface. In the cross-linked x-ray structure (pdb id: 2F5O) G:C base pair is in the interrogating position (Figure 5-2, G:C base pair is shown in blue color). We built a new structure base on the same x-ray structure, with the G:C base pair slided to the position of the next base pair on the left in the view of Figure 5-2. To build such a structure, we first elongated the left end of the duplex in Figure 5-2 by adding one nucleotide on each stand, and deleted one nucleotide from each stand on the right end. The new structure was saved as a pdb file. Then all atoms from the base groups were deleted from the pdb file, and the names of the nucleotides were changed to the match the original sequence of the DNA duplex. The missing bases were built using LEAP module in Amber. In this application

only the backbone atoms of DNA in this new built structure were used as target, the misplaced atoms of the base groups would not affect the simulation results.

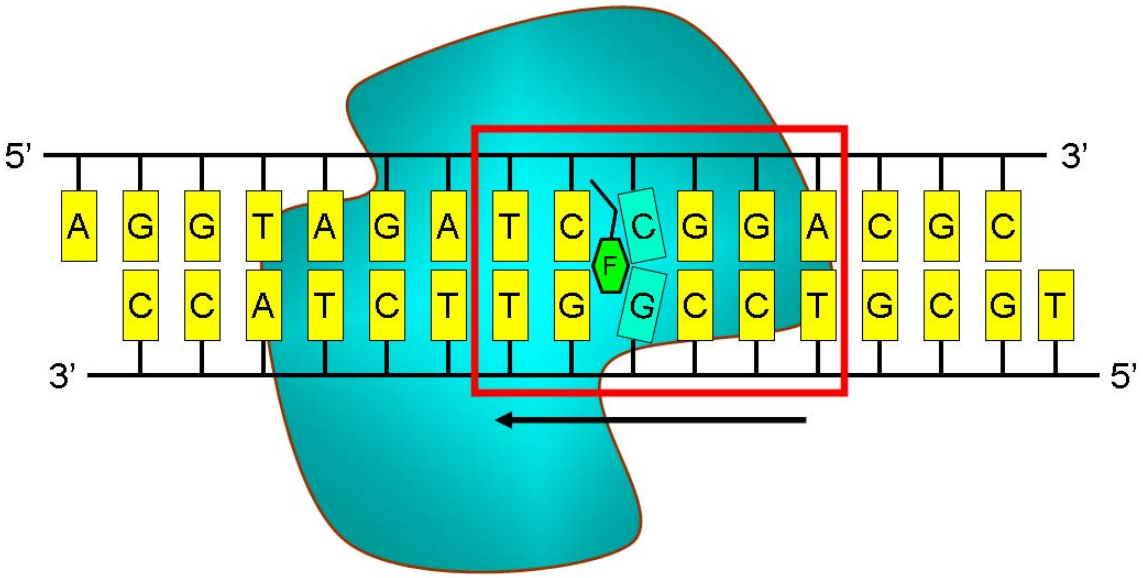

Figure 5-2. The scheme of the targeted MD for DNA sliding. The shape of the protein is shown in blue in the background. The DNA duplex and its base pairs are shown in black lines and yellow (or blue) blocks. The wedge residue Phe114 is shown in green block, which is inserted between two C:G base pairs. The buckled G:C base pair is shown in blue color. The targeting force is added on the six base pairs which are inside in the red block. Relative to the protein, the DNA duplex is forced to slide to the left in the current view.

Using this structure as the targeting structure and the x-ray structure as initial

structure, we can simulate a pathway between these two conformations. The force constant is 5 kcal/mol×Å^2. The force was applied on the atoms C5', O5, P, O1P, O2P, O3', C3', C4' of the six base pairs centered at G:C and new interrogated base pair. These six base pairs are circled in the red block in Figure 5-2. The targeting RMSD value was changed gradually from the original difference between starting structure and targetting structure to 0 over 500 ps simulation.

5.2.5 COM pseudo-dihedral angle

MacKerell *et al.* have been using the COM pseudo-dihedral angle and umbrella sampling method to measure the free energy profile of base flipping in solvent(*79-81*) and in protein/DNA complex(*82, 83*). The concept is shown in Figure 5-3. There are four sets of atoms defined for the pseudo-dihedral angle: (a) COM1, the guanine and cytosine bases forming a base-pair 5′ to the flipping guanine, (b) COM2, the sugar attached to the 5′ cytosine base, (c) COM3, the sugar attached to the flipping guanine base, and (d) COM4, the flipping guanine base itself.

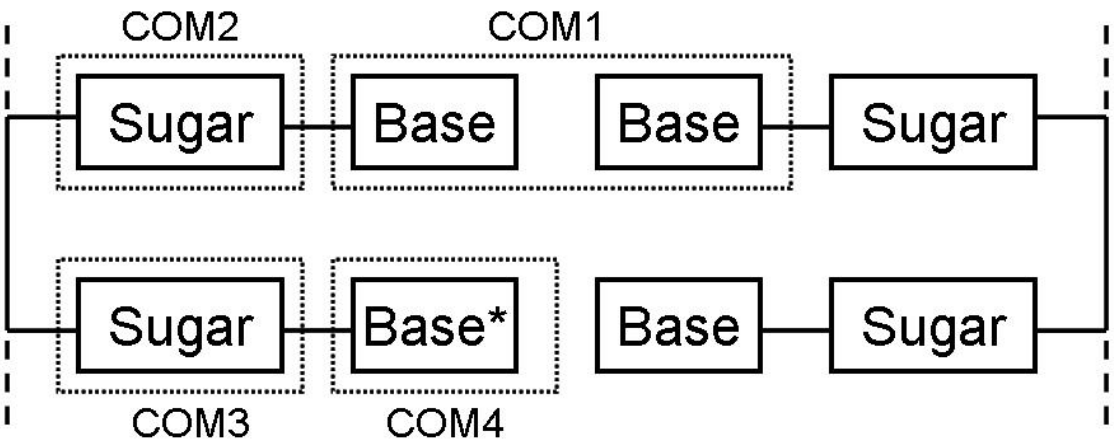

Figure 5-3. The scheme of MacKerell's COM pseudo-dihedral angle concept for the base flipping. The base group labeled "*" is the targeted base to be flipped.

In our analysis, we used this concept as the measurement of the base flipping in our normal MD simulations and targeted MD simulations. The sample script for calculating the pseudo-dihedral angle is shown in Appendix G – Pseudo-dihedral angle calcualtion.

5.3 Results and discussion

5.3.1 Targeted MD simulations for 8OG base flipping

DNA base flipping is a common theme in DNA base excision repair mechanism. Enzymes that would not otherwise be able to access the bases can perform modifications and alterations to the flipped-out bases. This is a relatively slow process. A recent kinetic study was done on Fpg/DNA complex using stop-flow fluorescence. The time-dependent

intrinsic tryptophane fluorescence changes were monitored. This study shows that 8OG everted from DNA duplex to the active site within 50 ms after the Fpg encountered 8OG in the DNA duplex(*169*). Studies on the methyltranferase and uracil DNA glycosylase also showed that the base eversion occurred at millisecond timescale (*170, 171*). This is still beyond the reach of today's molecular dynamics simulations that are typically limited to nanosecond time scale. For Fpg/DNA containing 8OG complex system, there are several x-ray structures available, including the structures with intact, intrahelical base pairs(*160*), and the structures with flipped, extrahelical base groups(*121, 160*), which are the start and the end of the base flipping process. The major differences of the 8OG in these two sets of structures are the COM pseudo-dihedral angle (intrahelical to extrahelical), and the glycosidic angle (*anti* to *syn*) of 8OG, along with other backbone changes.

Using the parameters and procedure described in the methods section, we performed the targeted MD simulation for 500 ps. The results are shown in Figure 5-4. The RMSD of 8OG, loop, and protein were calculated. The results are shown in panel 1 of Figure 5-4. Relative to the flipped 8OG structure, the RMSD of 8OG decreased from about 6.5 to 1.2 Å, showing a transition from an intact, intra-helical conformation to the flipped, pre-excision conformation. The stability of the binding loop has shown correlation with the presence of the substrate. The loop is ordered in all x-ray structures of Fpg/DNA complex with substrate (flipped damaged base) inside in the active site (*121, 152, 160*), and there is no electron density for the binding loop in the Fpg/DNA complex without substrate in the active site. There is no study showing the significance of the flexibility of

the binding loop in the base recognition and flipping. In this simulation, the initial conformation of the binding loop was built in based on the x-ray structure 1R2Y, in which 8OG is in the active site and the loop was present in the x-ray structure. It was also the reference structure for the RMSD calculation. In this simulation we see that the original conformation was stable for the first 300 ps, the RMSD value remaining below 2 Å. Then, when the flipping base was close to the active site (RMSD about 3 Å), the loop drifted away from the original position and the RMSD value fluctuated between 3 and 4 Å. After 8OG bound into the active site at 425 ps, the loop folded back to its original conformation rapidly. This is consistent with what is shown in x-ray structure. It is worth pointing out again that in this targeted MD simulation, additional force was only applied on to the 8OG group. It seems that the flexible binding loop can easily open an entrance for the 8OG base group to enter the binding site, and then the specific interactions will lead to tight binding for lesion recognition.

The pseudo-dihedral angle of 8OG's flipping (Phi) and its Chi angle were also calculated, the result is shown in the second panel of Figure 5-4. The results show that 8OG flipped out of the duplex first at about 360 ps, and then the glycosidic bond rotated into *syn* conformation at 425 ps.

In the x-ray structure with flipped 8OG(*121, 160*), 8OG forms four hydrogen bonds to Fpg: between 8OG N1 and T223 Oγ, between 8OG N7 and the backbone O in S219, between 8OG N2 and E77 Oε and a network of hydrogen bonds between 8OG N6 and the N atoms of residues 221 to 224 (Figure 3-9). In panel 3 the distance analysis shows the formation of these interactions. Comparing panel 2 and 3, we can clearly see the

correlation between the formations of the hydrogen bonds with the conformation change of 8OG. The hydrogen bond between 8OG N1 and T223 Oγ formed as soon as the base flipped (referring to the phi angle jumping at 360 ps in panel 2). The hydrogen bonds between 8OG N2 and E77 Oε and a network of hydrogen bonds between 8OG N6 and the N atoms of residues 221 to 224 were formed when 8OG changed its conformation from anti to syn at 425 ps (referring to the chi angle change in panel 2). The distance between 8OG N7 and the backbone O in S219 was still out of the range for hydrogen bonding, however, there was also a steep jump at the point where 8OG had *anti* to *syn* transition.

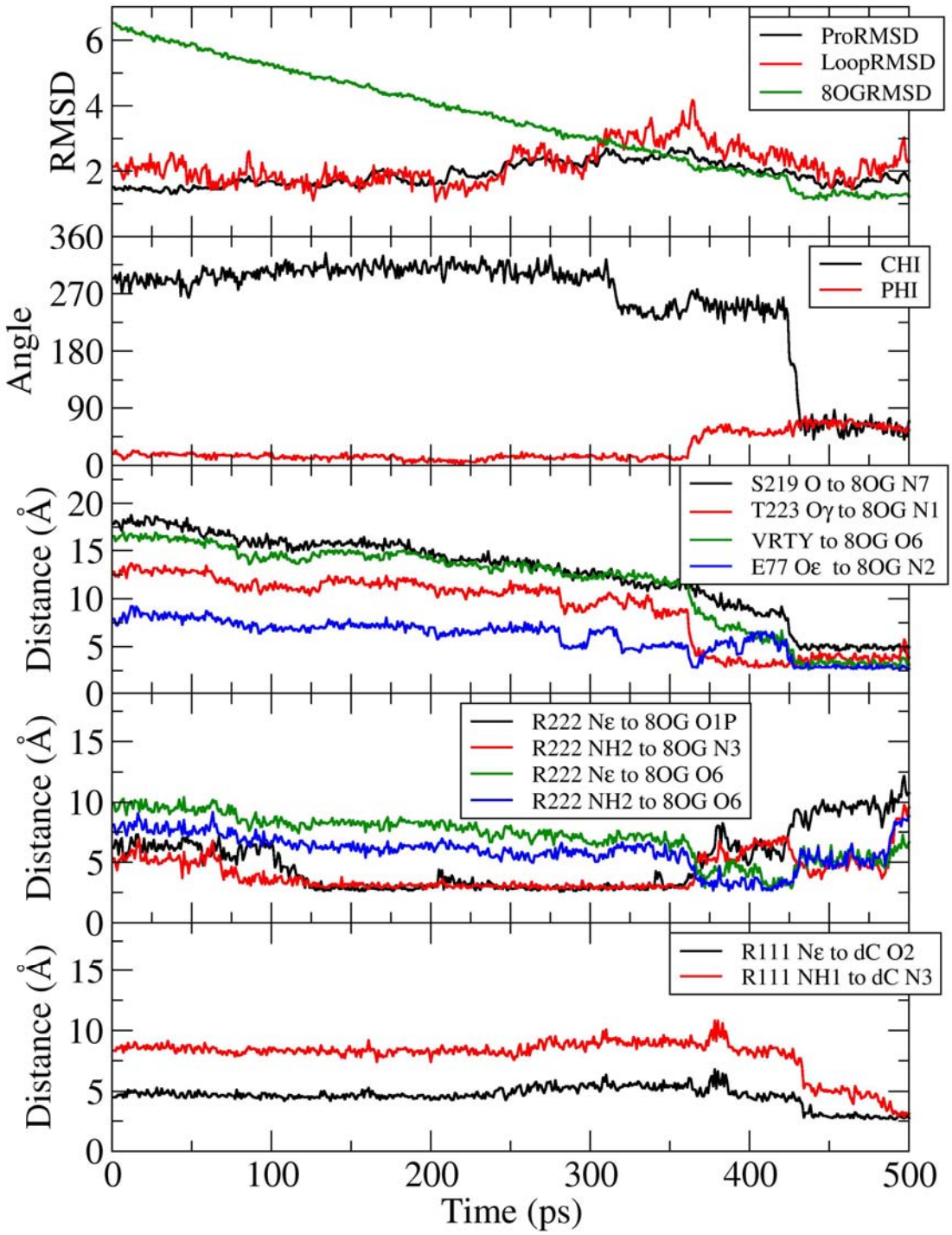

Figure 5-4. The analyses of the targeted MD trajectory. Panel 1 shows the RMSD of the protein, the binding loop, and the 8OG base group. All RMSD

values are calculated with fitting the structures with all Cα atoms of the protein. Panel 2 shows the chi angle (glycosidic angle) and phi angle (COM pseudo-dihedral angle) of 8OG. Panel 3 is the four hydrogen-bonds present in x-ray structure 1R2Y. Panel 4 shows the formation of the hydrogen bonds between R222 and 8OG. Panel 5 is the formation of the hydrogen bonds between R111 and the widowed cytosine.

Panal4 of Figure 5-4 shows the formation of interactions between R222 and 8OG. R222 belongs to the binding loop. In this trajectory it formed hydrogen bonds with O1P and N3 of 8OG in the middle of the transition, these interactions may help the process of base flipping. After 8OG entered the active site, there was no hydrogen bond between the sidechain of R222 and 8OG present, which is also consistent with what we have seen in the x-ray structures.

R111 is a strictly conserved residue. In all x-ray structures of Fpg/DNA complex containing the widowed cytosine, R111 always is inside of the duplex and forms two hydrogen bonds with cytosine. Panel 5 shows the spontaneous formation of these two hydrogen bonds. After 8OG flipped out from the intra-helical position, the hydrogen bond between R111 Nε and dC O2 formed first, and then the other hydrogen bond between R111 NH1 and dC N3 also formed before the simulation ended. This phenomenon is especially significant because there was no additional force applying on either of these two residues.

5.3.2 Phe114 reinsertion in targeted MD simulations of DNA sliding

We also applied targeted MD to simulate the DNA sliding. Using the method described in the methods section, we created a structure with slided Fpg/DNA structure. The additional force was applied to force DNA double strands to slide. After 1 ns of targeted MD simulation, we resumed it with a free MD simulation starting from the final structure of the targeted MD simulation. The analysis results of these two nanoseconds are shown in Figure 5-5.

In the recently solved x-ray structures of Fpg/DNA with intact interrogated base pairs, the aromatic ring of Phe114 is partially inserted inside of the duplex(*160*), and DNA cannot slide without clashing with Phe114. In this simulation, we first monitored the conformational change of Phe114 along with the progress of the DNA sliding. The reference structure is the conformation of Phe114 in the x-ray structure 2F5O. We overlapped the structure of the protein, and then calculated the RMSD between the conformation of Phe114 in the reference structure and that of each frame. The results are shown in panel 1 of Figure 5-5. The RMSD value increased from about 1 Å to about 6 Å at the end of 1 ns. After about 1120 ps of normal MD simulation, the RMSD value of Phe114 decreased dramatically from about 6 Å to 2 Å, and fluctuated between 2 Å and 4 Å in the rest of the simulation.

To understand the cause of the sudden RMSD decrease at 1120 ps, we did two further analyses. One analysis is the position of Phe114. We measure the distance between the mass center of the aromatic ring of Phe114 to the mass center of the two

base pairs which surrounded Phe114 before the sliding (initial pocket) and after sliding (final pocket). The results are shown in panel 2 of Figure 5-5. At the beginning of the simulation, the sidechain of Phe114 was partially inside of the duplex, with the distance being below 4 Å. The distance to the final pocket was about 6 Å. At the end of the targeted MD simulation (1000 ps), Phe114 was outside of both pockets. At 1120 ps, the distance to the final pocket suddenly decreased to about 3.5 Å, while the distance to the original pocket was at about 5 Å. These data, along with our visual analysis of the simulation trajectory, shows that the aromatic ring of Phe114 left the original pocket, and entered the final pocket created by the DNA sliding.

To exhibit the conformation change of the sidechain of Phe114 along with the simulation, we also measured the chi1 and chi2 angles, the two torsional angles of the sidechain of Phe114. The results are shown in panel 3 of Figure 5-5. At 500 ps when the Phe114 was completely out of the initial pocket (the distance to the initial pocket was 4.5 Å, this value was obtained by visually analyzing the sliding trajectory.), the values of both torsional angles had significant change. The chi1 angle had changed from -90° to -160°. The chi2 angle had changed from 40° to 140°. Soon after the chi2 angle reached 160°, there was a ring flipping, which shown as chi2 angle changing from 150° to -50°. There was no significant change for both angles in the rest of the targeted MD simulation. In the normal MD simulation at 1120 ps, when the aromatic ring of Phe114 inserted into the final pocket (panel 2 of Figure 5-5), both angles switched back to their original values. This indicates that in this simulation the wedge inserted into the second pocket in the same way as it did in the initial pocket.

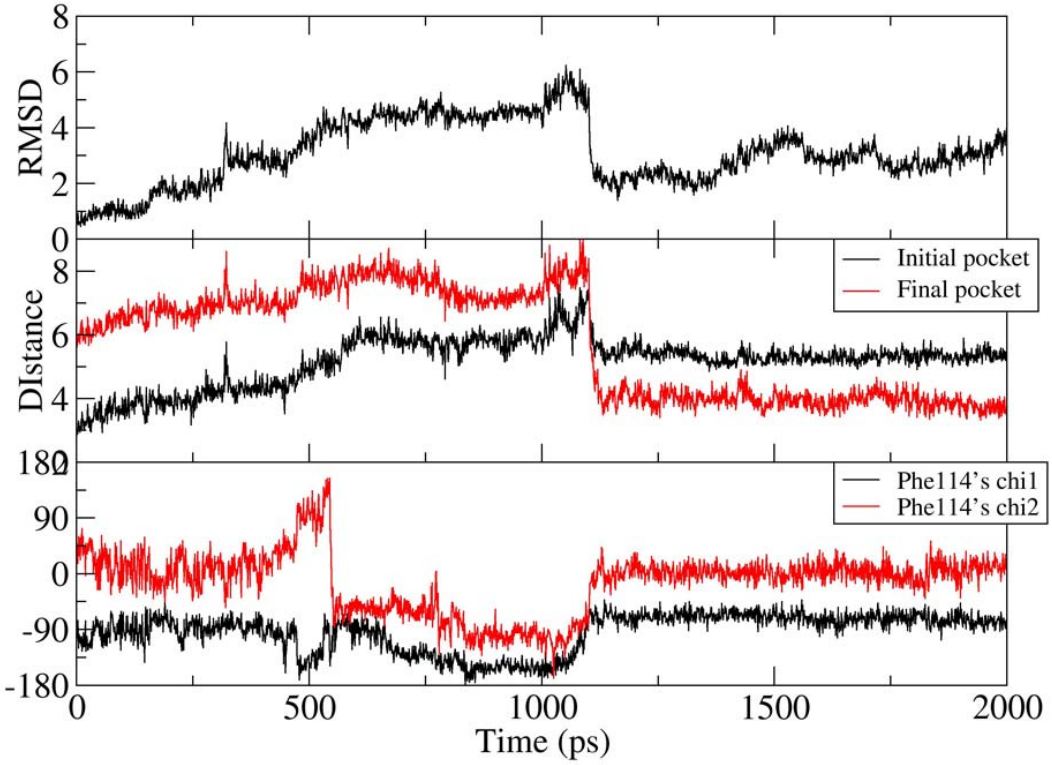

Figure 5-5. The analyses of the targeted MD trajectory (0~1ns) and extended free MD trajectory (1~2ns). RMSD: using all heavy atoms in Phe114, with fitting on the protein. Distance: The distance between the mass center of Phe114's ring and the two base pairs around it. Chi1, Chi2: The angles of Phe114's side-chain.

We compared the final structure from the 2000 ps simulation with the original structure built based on the x-ray structure 2F5O. Figure 5-6 shows the overlap of the

four centeral base groups in the final structure (colored by atom type) and in the x-ray structure (in green). The new pocket formed by the two base pairs has a similar shape as the initial pocket. In the view of Figure 5-6, the base pair above the wedge Phe114 is buckled. The conformation of Phe114 in the final pocket is approximately the same as its conformation in the initial pocket.

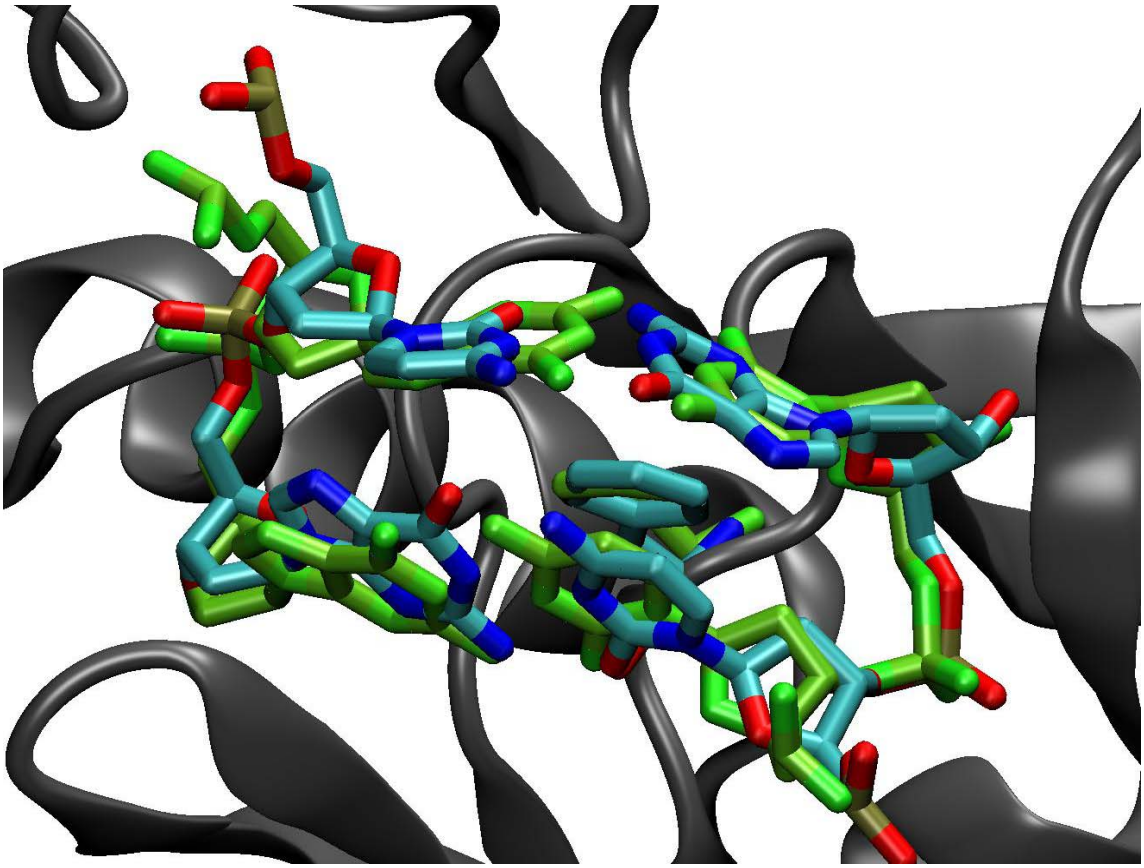

Figure 5-6. The overlap of the central four base groups in the final structure from extended MD simulation and in the original x-ray structure. The structure colored by atom types is the simulated structure. The one in green is the x-ray

cross-link structure. The wedge residue Phe114 is also shown.

We also extended the targeted MD simulation using restrained MD simulation. In the extended simulation, the initial structure was the final structure from the previous targeted MD simulation (the structure at 1ns in Figure 5-4). The same restrain mask and restrain force constant as targeted MD were used, with the target RMSD value being 0. The results are shown in Figure 5-7. In the free MD simulation, the RMSD of Phe114 went up to 5.5 Å and stayed for about 100 ps, and then decreased rapidly to below 2.0 Å. Along with RMSD value decrease, the distance between Phe114 and the second pocket decreased from 9.0 Å to 3.5 Å. In this restrained MD simulation, Panel 1 in Figure 5-7. shows the RMSD value all heavy atoms in Phe114, with fitting on the CA atoms of the protein. The RMSD values of Phe114 had no change in 2 ns simulation, which is opposite to the rapid changes in 100 ps in free MD simulations. The distances from Phe114 to the initial and final pocket, and the two dihedral angles of Phe114's sidechain had very little deviation from the starting point in the 2 ns simulations. This shows that the flexibility of the backbone of DNA is important for the new intercalation of the wedge residue Phe114 to the new pocket during enzyme/DNA translocation.

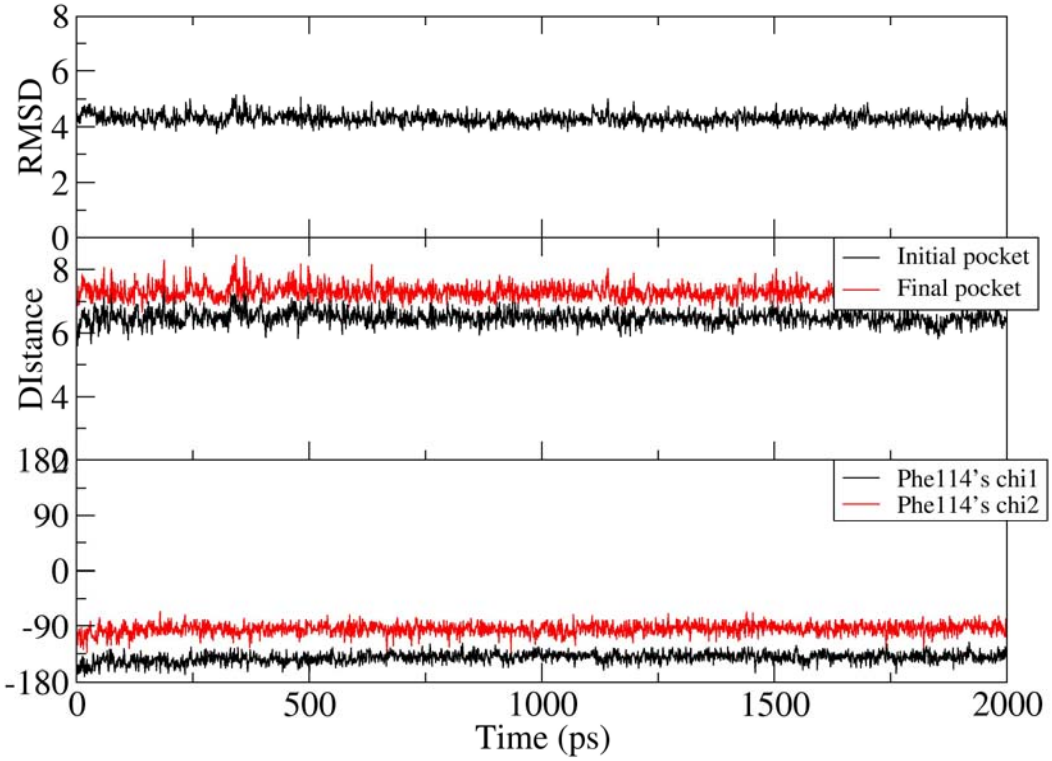

Figure 5-7. The analyses of the extended restrained MD trajectory. The initial structure is the final structure from targeted MD simulation (the structure at 1ns in Figure 5-4). The same restrain as targeted MD was used, with the target RMSD value being 0. Panel 1, RMSD: using all heavy atoms in Phe114, with fitting on the protein. Panel 2, Distance: The distance between the mass center of Phe114's ring and the two base pairs around it. Panel 3, Chi1, Chi2: The dihedral angles of Phe114's side-chain.

5.3.3 Long normal MD simulations

Besides the targeted MD simulations, we also carried out long normal MD simulation to explore the evolution of the conformation of the complex without additional forces. They were labeled as VIC3 (the structure built based on x-ray structure 2F5O), VOIC3 (the same as VIC3 except that the guanine in the interrogating site was changed to 8OG), F114AGC (the same as VIC3 except Phe114 was mutated to alanine), and F114AOGC (the same as F114AGC except that the guanine in the interrogating site was changed to 8OG). For each system, two simulations were carried out from the same starting structures, named as A and B. As mentioned in the methods section, eight simulations were performed, and each individual simulation is more than 200 ns long.

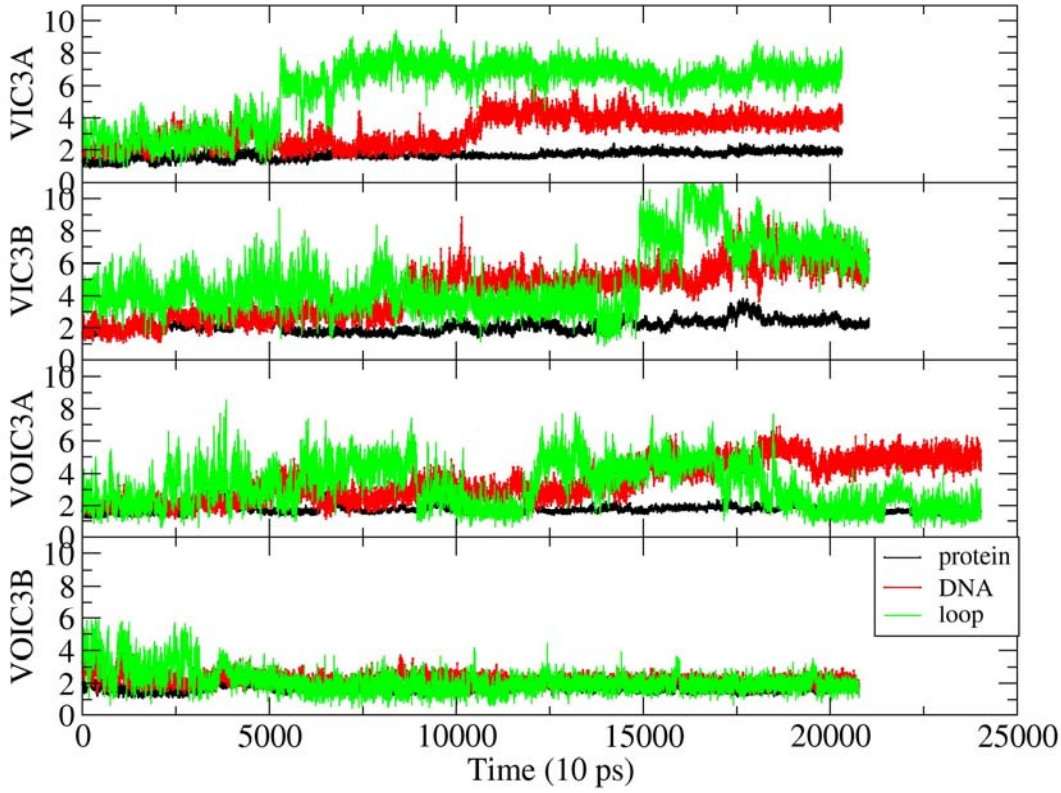

Figure 5-8. The RMSDs of long MD simulations on WT Fpg with G:C and OG:C base pairs. .

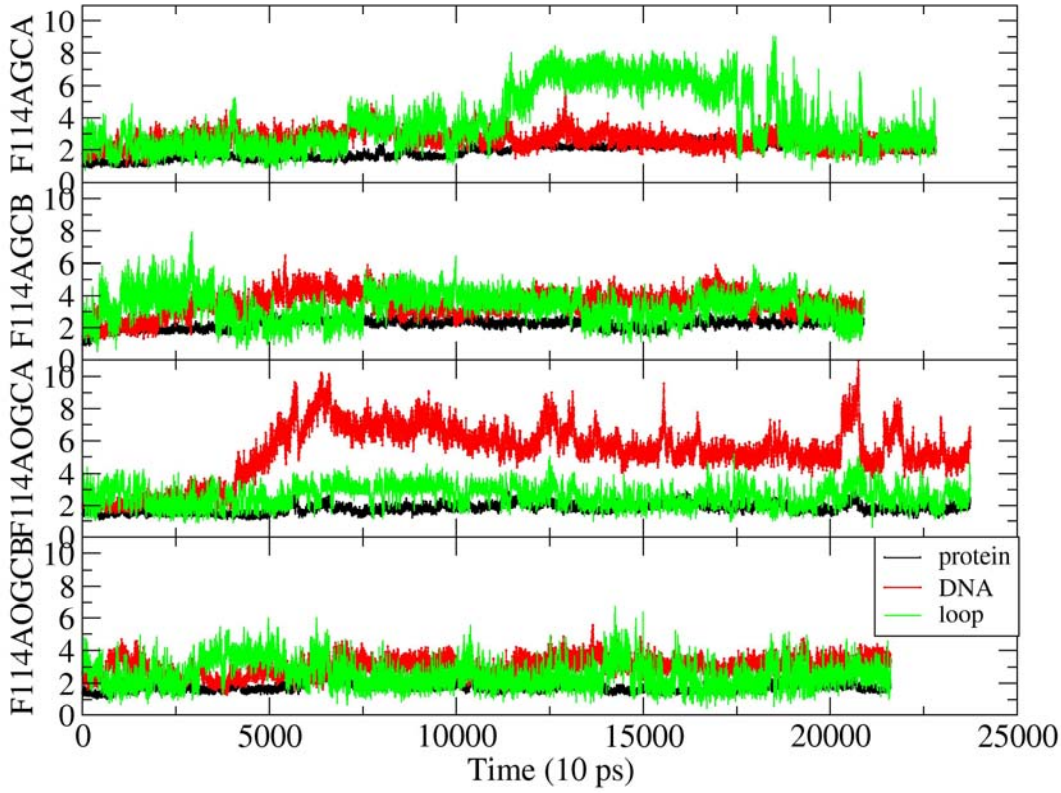

Figure 5-9. The RMSDs of long MD simulations on F114A Fpg with G:C and OG:C base pairs.

To analyze the stability of the structures under the simulation conditions, the RMSDs of protein Cα atoms, lesion site residues (8OG:C or G:C and flanking base pairs) and the βFα10 loop were computed for each of the eight simulations. Proteins in all simulations were quite stable. In all simulations, the RMSDs of proteins remained below 2 Å.

The structures of the DNA were stable in the first 20 ns in all simulations. The RMSD value started increasing in several of the simulations. The first

RMSD increase occurred in the simulation F114AOGCA. We visually analyzed this trajectory and observed that the DNA sliding occurred in the simulation.

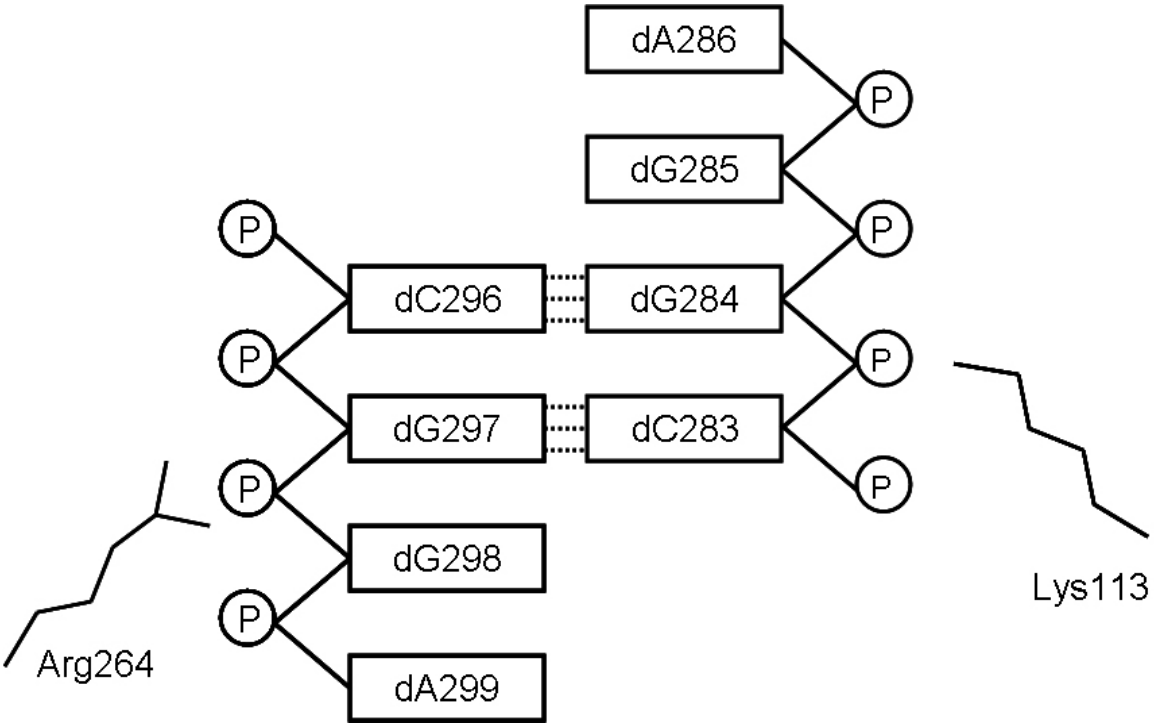

Figure 5-10. The illustration of the DNA double strand in the interrogating site and the neighboring residues Arg264 and Lys113. To clarify the representation, only 8 nucleotides are shown.

Figure 5-10 shows a scheme of the fragment of DNA double strand near the interrogating site and the two neighboring residues interacting with the phosphates. R264 is close to the strand containing dG, and K113 is close to the strand containing dC. This

is the conformation before the sliding occurs. We measured the distance from R264 CZ or K113 NZ to the P atom of the nucleotides. This analysis was only done in the first 90 ns of the whole simulation in which the translocation occured. The results are shown in Figure 5-11.

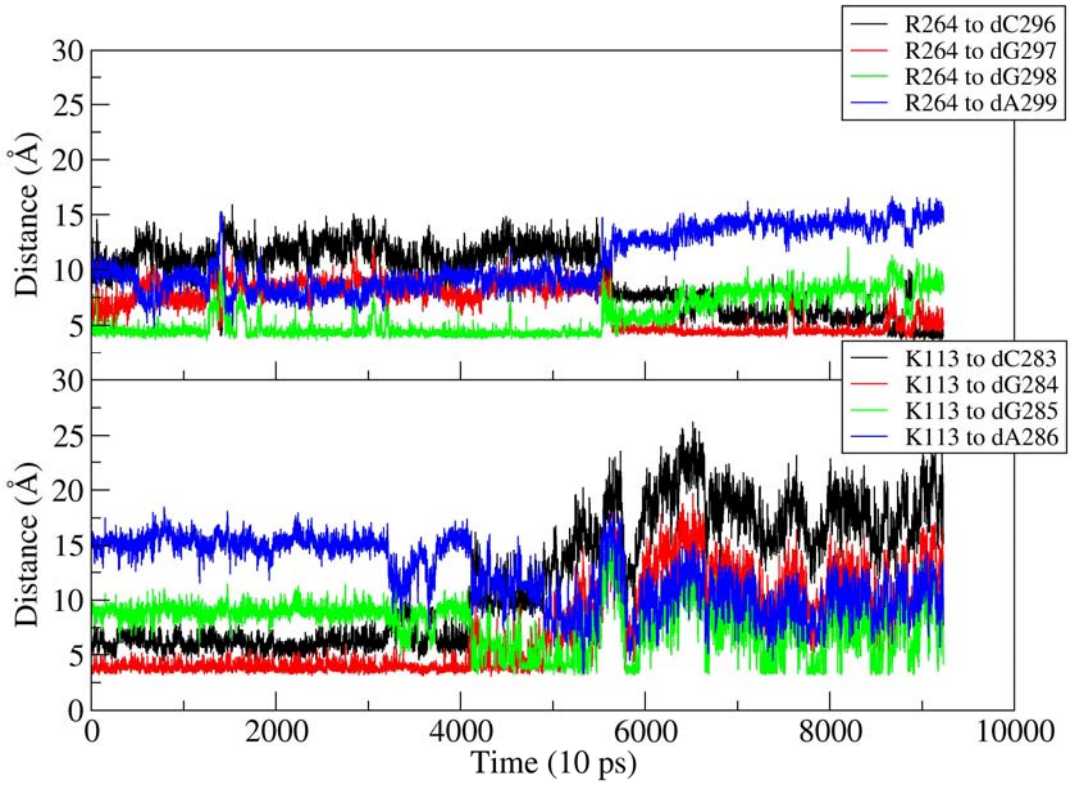

Figure 5-11. The two important residues in the sliding. The distance is defined by the R264's CZ or K113's NZ to the P atom of the nucleotides.

The upper panel of Figure 5-11 shows the position of R264. It was hydrogen-bonding with the phosphate of dG298 at the beginning of the simulation. At 56 ns when the DNA sliding occurred, this distance increased from 4 Å to 7 Å. On the other hand, due to the sliding, nucleotide dG297 came close to the residue R264. The distance between these two groups decreased from 8 Å to 4 Å.

The story for the interactions between the K113 and the strand containing dC is more or less the same. In the initial structure, K113 interacted with dG284. The interaction only lasted about 48 ns. Staring at 40 ns, K113 also formed interaction with dG285 from time to time. This interaction lasted till the end of the simulation. Interestingly, the interactions involving K113 became more flexible after 56 ns when the DNA sliding occurred. In this simulation the wedge residue F114 was mutated to Alanine. Therefore there was no stacking interaction between the wedge and the neighboring base pairs to stabilize the complex structure. This could be the reason why the interactions between K113 and the DNA strand containing dC were less stable.

F114 is an important residue which has been shown to act as a wedge inside the intact, intrahelical DNA base pairs(*160*). We monitored the position of the aromatic ring of F114 relative to the base pairs. The way we measured the position is shown in Figure 5-12. In the initial structure, the sidechain of F114 was between base pair G297:C283 and G298:C282. The mass center of these four base groups was defined as position 0. After the sliding, F114 can be inserted into the pockets above or below position 0. The position -1 and position +1 are the new positions F114 can relocate after sliding over one base pair, which are defined as the mass center of the four corresponding base groups. We

measured the distance between the mass center of F114's aromatic ring and these three positions, and the results are shown in Figure 5-13.

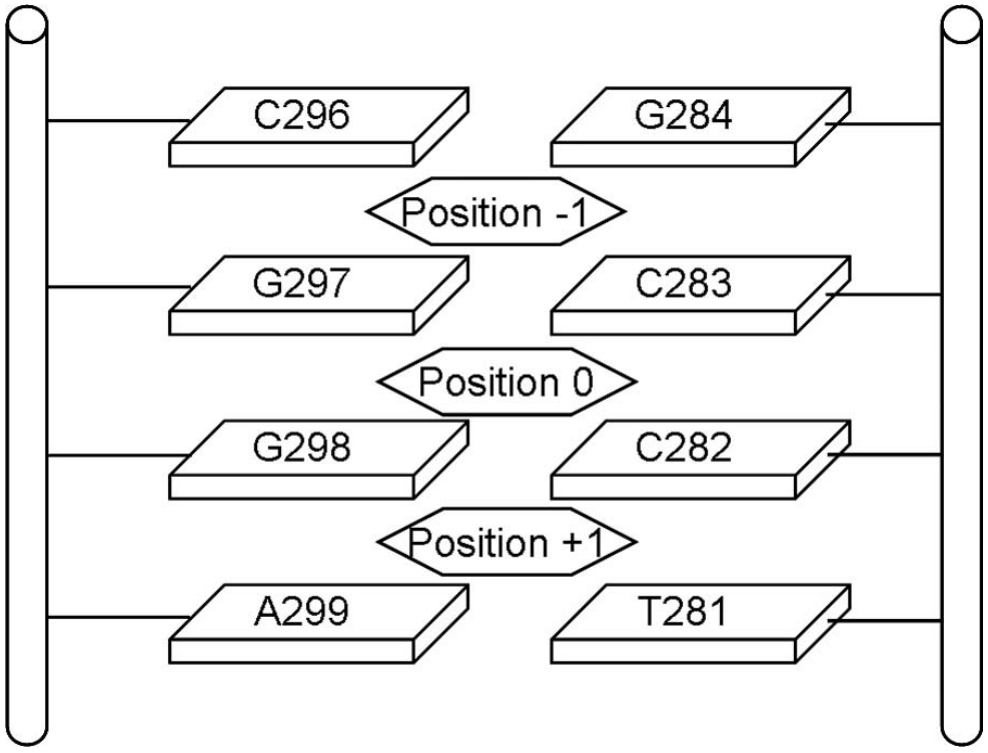

Figure 5-12. The original position of wedge (0) and two potential positions after sliding (-1 and +1). Each position is defined as the mass center of the four surrounding base groups.

In both wt Fpg/DNA containing G:C simulations (VIC3A and VIC3B), the wedge left the position 0,evident as the distances to position 0 were increased over 4.5 Å. By visual examination we found that at this distance the aromatic ring of F114 is outside of

duplex and free to rotate. In the simulation VIC3B, the distance to position +1 became shorter than to the position 0 for the last part of the simulation, which is the signature for the occurrence of sliding. However, due to the flexible C-strand, the new insertion did not go as deep as in the disWG0.

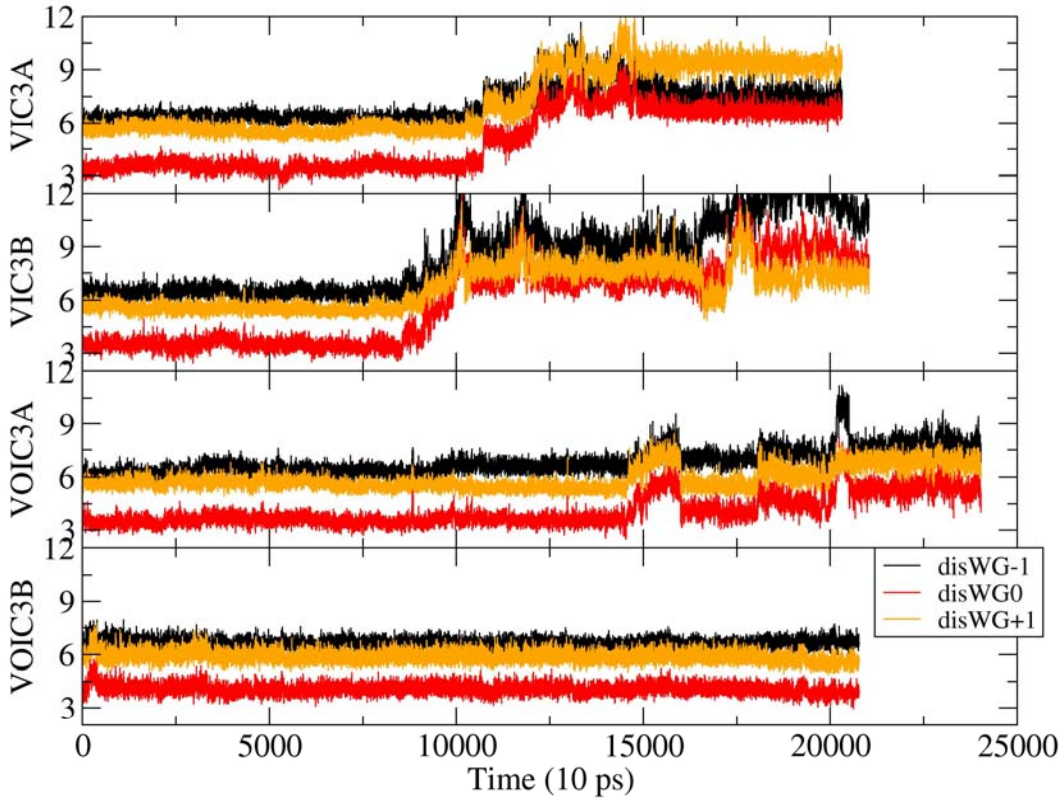

Figure 5-13. The wedge's positions in the simulations of wt Fpg. disWG0 is the distance between the mass center of the aromatic ring of F114 to the position 0, disWG+1 is for that of the position+1. And disWG-1 is for that of the position -1. These three positions are defined in Figure 5-12.

In the simulations of wt Fpg/DNA containing 8OG:C (VOIC3A and VOIC3B), there was no large increase in distance. In simulation VOIC3B, all three distances were maintained all along the 210 ns simulation. In VOIC3A, the distances had been fluctuated near 205 ns. We examined the structures and found that the increase in distance at this point was not because of the DNA sliding, but because of 8OG:C base pair breaking. The structure of the breaking base pair (Figure 5-14) and more extended simulations are shown in the next section.

5.3.4 Spontaneous 8OG:C base pair breaking and the intermediate state in the base flipping

In one fo the long MD simiulations (VOIC3A) we observed 8OG:C base pair breaking. The structure is shown in Figure 5-14. The hydrogen bonds between 8OG and its partner cytosine are broken. 8OG partially flipped out of the duplex. The upstream and downstream base pairs are intact. The pseudo-dihedral angle of 8OG is 82º (The definition of the pseudo-dihedral angle is in the method section 5.2.5).

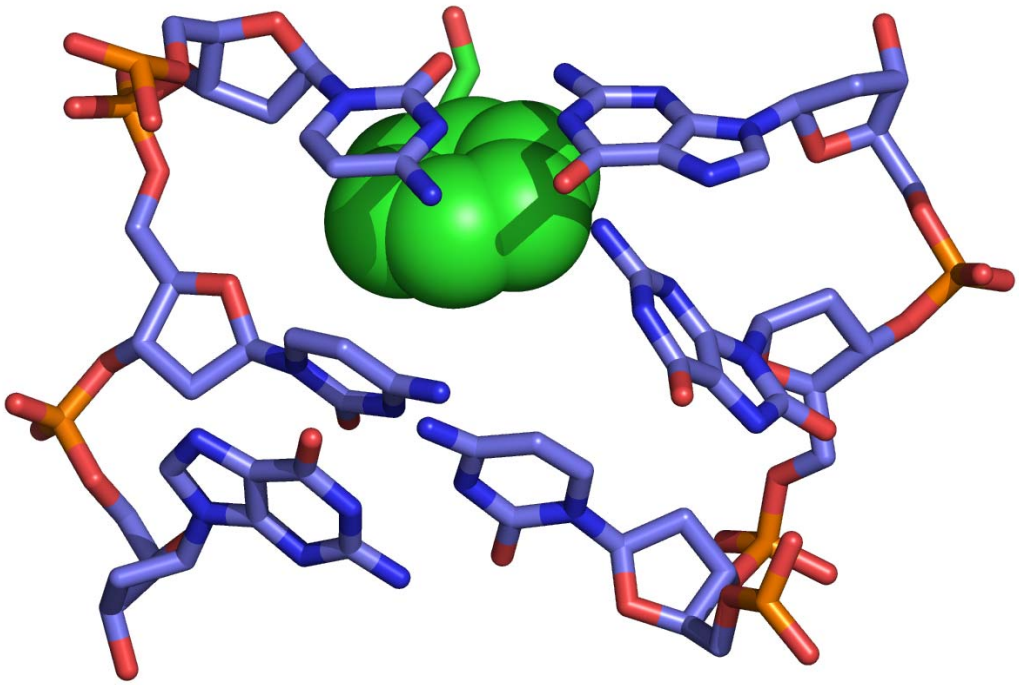

Figure 5-14. The broken base pair in the VOIC3A. In middle base pair is C:8OG. The base group on the right is 8OG. The wedge residue Phe114 is shown in green.

Recently a semi-open G:C base pair in the enzyme/DNA complex environment has been published(*162*). In that study Banerjee and Verdine used disulfide cross-linking technology and captured hOGG1, the functional analog of Fpg in human, examining a undamaged G:C base pair which is adjacent to an OG:C base pair. In the resulting x-ray structure, a guanine was extruded from the DNA duplex because of the disulfide cross-link formed between its partner cytosine and a cystine in the protein. The hydrogen bonds between this G:C base pair were broken. However, due to the effect of the adjacent 8OG, the guanine still remained in the major groove of the DNA. The authors

concluded that this semi-open structure is analogous to the early intermediate stage of the base flipping procedure.

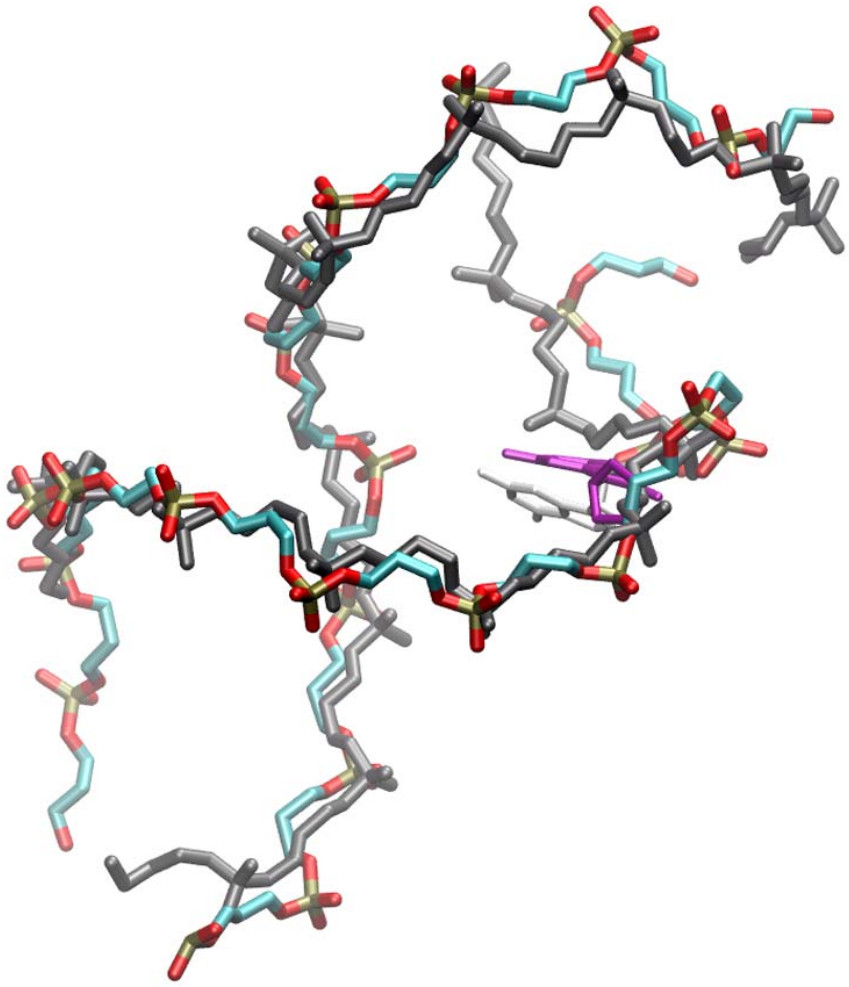

Figure 5-15. The backbone phosphate atoms overlap of the semi-open snapshot from our simulation (colored by atom type) with x-ray structure 2I5W (in gray). The flipped base in the simulated structure (8OG) is in purple, and the one in x-ray (dG) is white.

We overlapped the snapshot of semi-open 8OG with the recently published x-ray structure 2I5W. Figure 5-15 shows the backbone overlap of the two structures. We can see that the C-strands overlap with each other very well. The strand containing 8OG can be divided into two parts. The part on 8OG's 5' side overlap very well too. The one on 8OG's 3' side doesn't match each other closely.

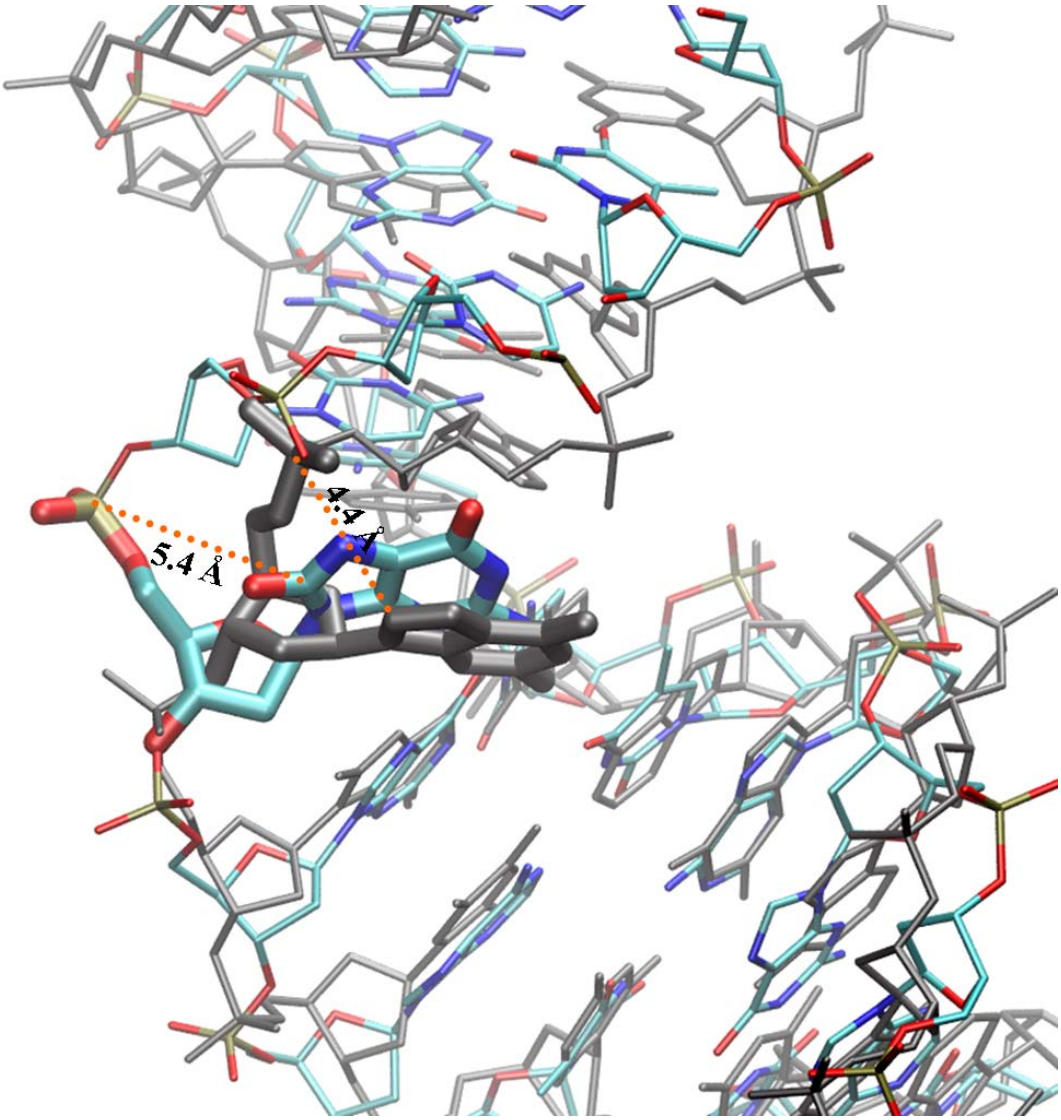

Figure 5-16. The closer look of the base-opening site. The structures are overlapped the same way as in Figure 5-15.The semi-open snapshot from our simulation is colored by atom type. The x-ray structure 2I5W is colored in gray. The flipping base in the simulated structure is 8OG, and it is undamaged guanine in the x-ray structure. The distance of the atom C8 and the phosphorus atom of the flipped nucleotide is measured. The distance in the simulated structure is 5.4 Å. The distance in the x-ray structure is 4.4 Å.

Figure 5-16 shows the closer look of the overlap. The two base groups have very similar locations. However, the positions of the phosphates of the semi-open nucleotides are very different. One possible explanation is the interaction between 8OG's O8 and the oxygens in the phosphate in the simulated structure. In the x-ray structure, the flipped nucleotide is an undamaged guanine. The phosphate group is closer to the nucleotide. In the simulated structure, the flipped nucleotide is 8OG. The O8 atom of 8OG has repulsive interaction with the oxygen atoms in its phosphate group. We measured the distance between the C8 atom and the phosphorous atom of the flipped nucleotide. This distance is 4.4 Å in the x-ray structure. It is 5.4 Å in the simulated structure.

According to the proposal of Banerjee and Verdine, this semi-open form is one early intermediate state in the base flipping process. To test the feasibility of the base flipping proceeding from this conformation, 20 simulations were started from this conformaiton. For these trajectories, the COM dihedral angles (defined in Figure 5-3) have been calculated. The results are shown in Figure 5-17.

We can see that in most of our simulations the COM angles are stable at the original value (~80 degree). Some of them have populations at about 30 degree.

From our simulations we also found a semi-open structure which is similar to the one Verdine *et al.*'s x-ray structure of hOGG1/DNA complex. However, further testing is needed to be done to test if this state is on the pathway of DNA base flipping or just a

dead end of a minor pathway.

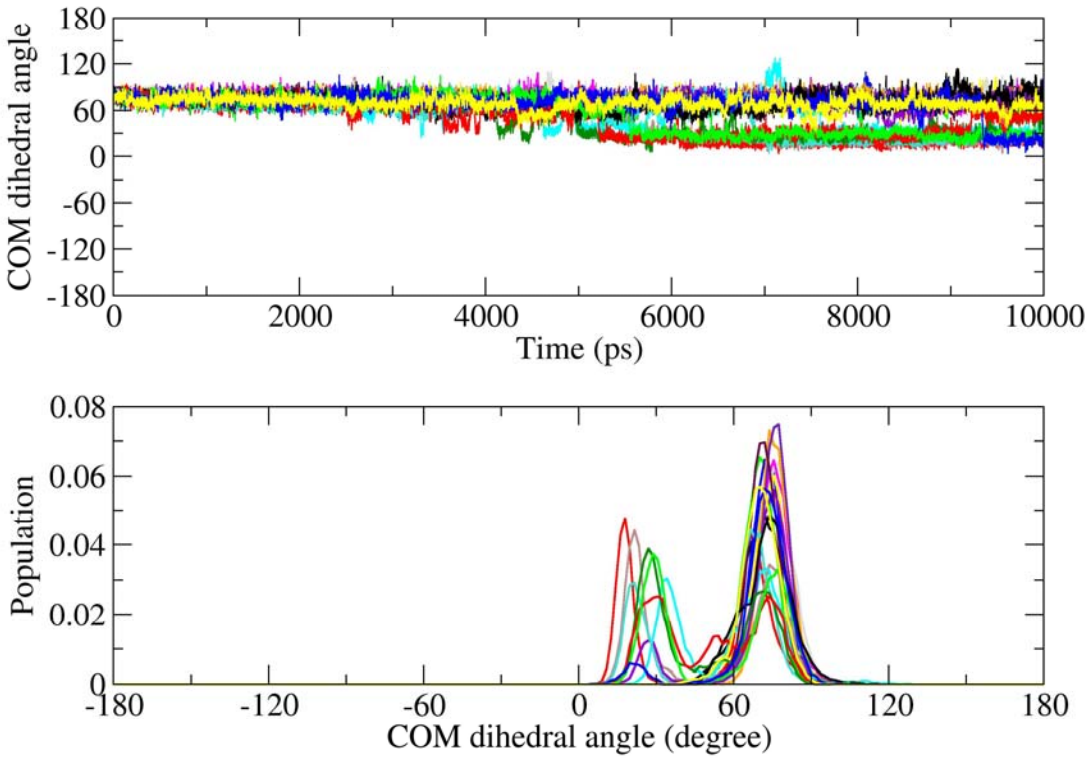

Figure 5-17. The COM dihedral angles from 20 independent simulations. The top panel shows the time sequence of the result. The lower panel shows the histogram result.

5.3.5 Two slow motion modes programmed in the topology of Fpg

Recent structural and dynamic studies on proteins indicate that static

structure along is not able to fully explain the mechanism of the protein functions. Normal mode analysis (NMA)(*172*) has emerged in recent years as a useful tool to elucidate the structure-encoded dynamics of proteins(*173*). An important conclusion drawn from those studies is that protein structures have evolved in such a way that their intrinsic structural flexibility, which is a combination of the normal modes, facilitates the functionally important conformational variations(*174*). In this study we used ANM(*164, 165*), a special type of NMA, which considers that the protein in the folded state is equivalent to a three-dimensional elastic network. The analysis was done on the ANM web server (http://ignmtest.ccbb.pitt.edu/cgi-bin/anm/anm1.cgi)(*175*).

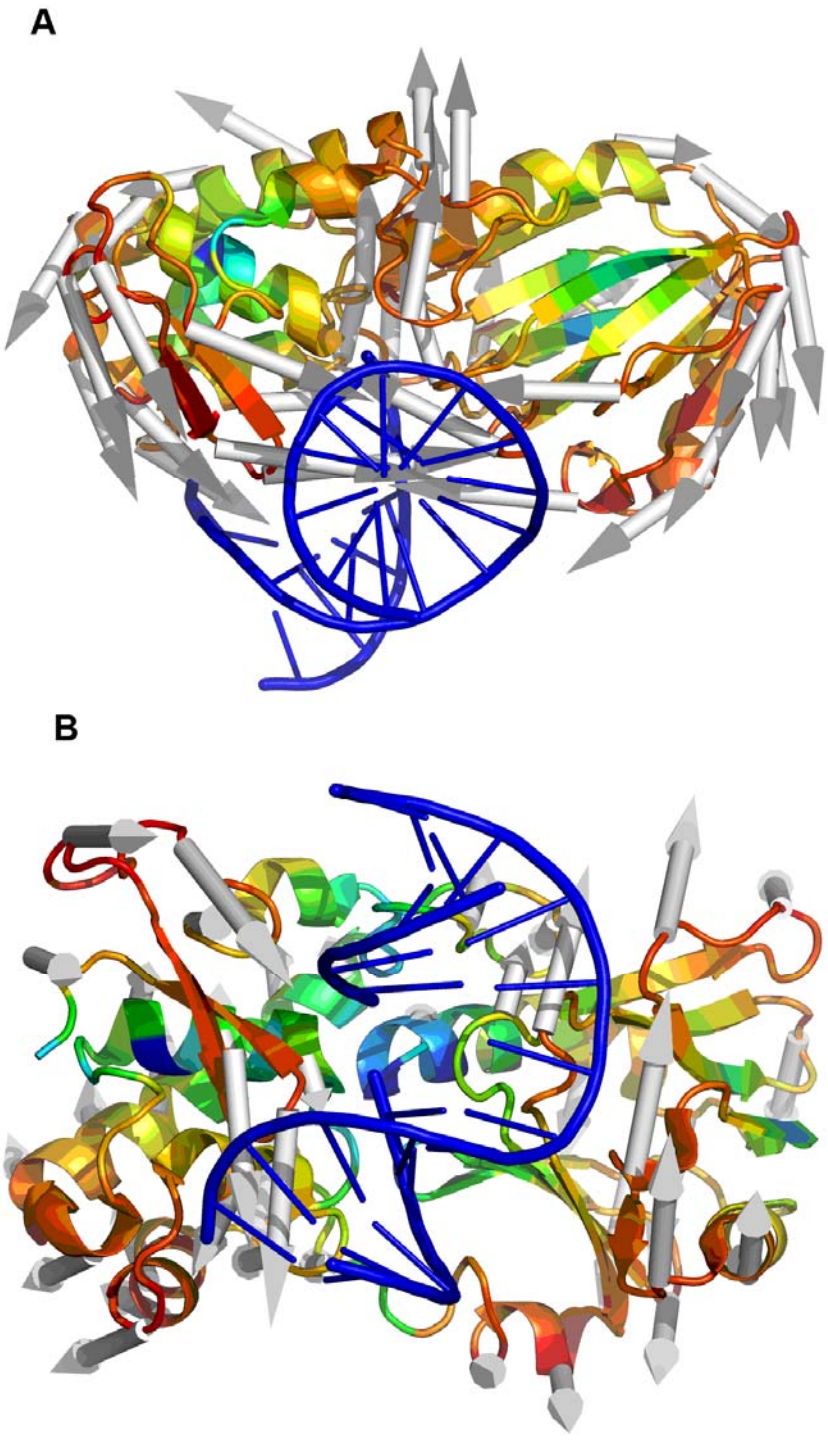

Figure 5-18. The first two slow motion modes of Fpg generated from ANM analysis. A is the first normal mode, bending mode. B is the second mode,

twisting mode. The cartoon structure of the protein is color coded by the magnitude of the vibrations. The magnitude increases from blue to red. The arrows show the directions of the motions. The DNA structure, which is not included in ANM analysis, was modeled in to show the relative positions. For easier visualizing, the mode A is shown from the top view, and the mode B is shown in the front view.

Figure 5-18 shows the two dominant modes, bending and twisting. Both of them are domain motions. The bending mode (Figure 5-18A) shows the motions of the two domains rotating in opposite directions around the hinge region between the two domains. The twisting mode shows the rotations of the two domains around the axis perpendicular to the hinge.

To explain the exact relation between these two modes and the function of Fpg still needs further investigation. Here we propose a hypothesis. The primary function of Fpg is sliding on the DNA duplex and searching for lesions. In the bending mode, when the two domains open up, there is more space for DNA sliding to process. When the two domains are close together, the sliding will temporarily stall to let the Fpg examine whether the base is damaged. The twisting mode is more related to the translocation. Imagine the DNA duplex as a long rope and Fpg as a mountain climber. The two domains, which have multiple interactions with each DNA strand, are like two arms of the climber. The twisting mode is like the motions of the two arms when the climber is climbing. The two arms reach out one after the other to grab the rope, and then pull the

body up to the forward direction. In Figure 5-10 and Figure 5-11 the two residues from both domains showed similar motion patterns in our regular MD simulations.

5.4 Conclusion

DNA recognition is proposed as a complex, multi-step procedure, in which each step plays a certain role in lesion discrimination(*30*). DNA sliding and base flipping are two essential steps in DNA lesion searching and recognition. However, due to the time scale of the DNA sliding(*176*) and base flipping(*170, 171*), this process cannot be simulated by using regular MD simulations which are in nanosecond time scale. In this study we used the targeted MD and long MD simulations (1.6 ms in total) were also performed.

R111 and F114 are two key residues for stabilizing the 8OG-flipped structure. F114 has also been observed to similar function in the intact intrahelical structures. In the base flipping simulations, R111 spontaneously entered the cavity created by the base flipping and formed hydrogen bonds with widowed cytosine. In the DNA sliding simulations, F114 left the original position and entered the new position along with the DNA sliding. Both final structures are consistent with x-ray structures. In the long MD simulations, we observed the 8OG:C base pair breaking. Structural overlap showed that this semi-open DNA structure is very similar to the recent x-ray structure in which hOGG1 binds to the DNA. Multiple simulations starting form this intermediate state showed that this is a stable conformation.

Chapter 6 The Role of Phe114 in Fpg in Searching and distinguishing Damaged DNA Base 8OG

6.1 Introduction

Efficient and accurate repairing DNA damage is essential to all living organisms (*6*). 8oxo-guanine (8OG) is one of the most common form of the oxidative DNA damage(*177*). Due to its ability to form Hoogstein-type base pair with adenine(*178*), failure to repair this lesion will cause G:C to A:T transversion(*179, 180*). In *E. coli.* and many other bacteria, it is rapidly repaired by DNA repair enzyme Fpg when 8OG pairs with cytosine(*19*). The damage recognition mechanism has not been fully understood yet. A series of x-ray structures solved recently exhibited the structures of the DNA/Fpg complex with normal G:C and A:T base pair at the interrogating site(*160*). In these structures, the targeted base pairs are intact, with the sidechain of Phe114 partially inserted inside of the duplex (Figure 6-1). In this particular research we are mainly interested in the role of Phe114 in searching the damage along with DNA sliding on the protein/DNA interface.

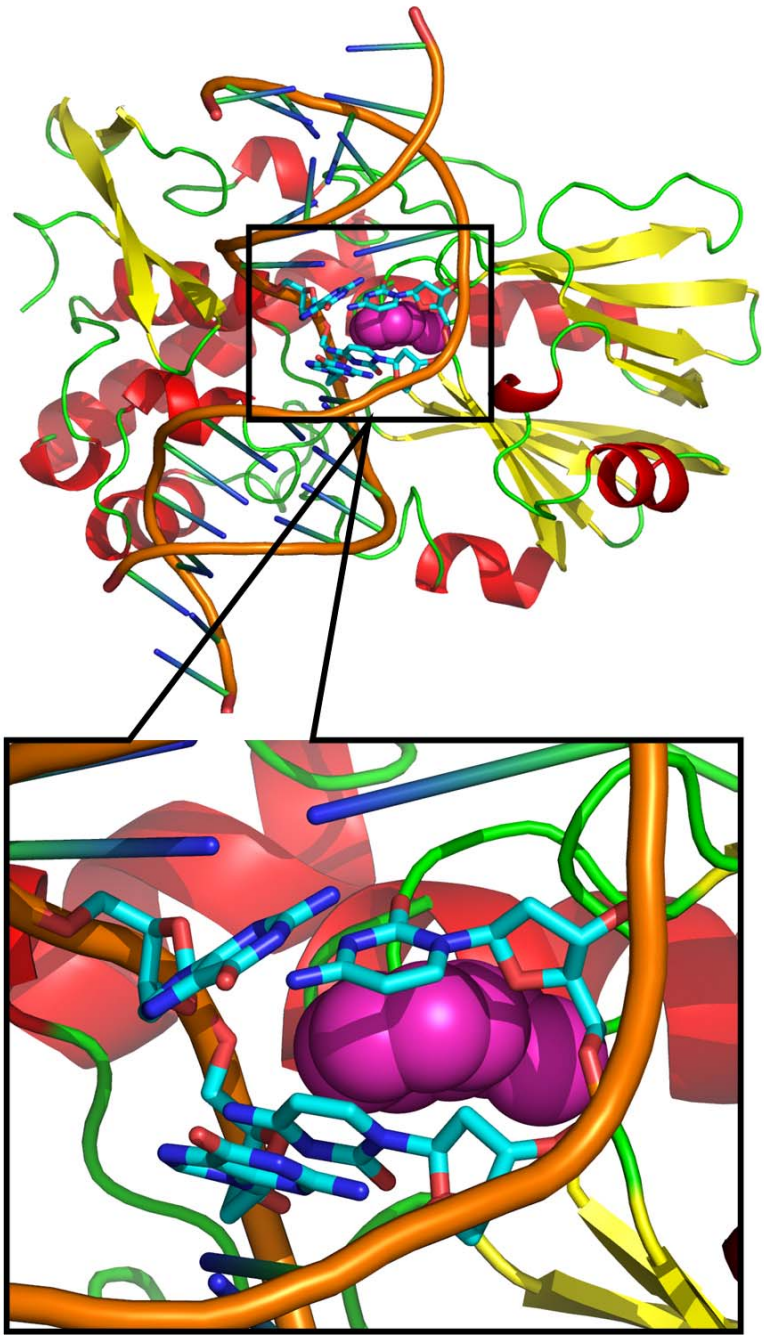

Figure 6-1. Structure of *B. st*. Fpg crosslinked to G:C containing DNA (pdb code: 2F5O). Atomic detail is shown for the DNA duplex and the protein is represented with a cartoon diagram. The wedge residue Phe114 is shown in sphere model.

There are several observations suggesting that Phe114 plays an essential role in damage recognition. First, among the known sequences of Fpg in different bacteria, this residue is strictly conserved. Second, in all available x-ray structure of Fpg/DNA complex, the aromatic ring of Phe114 is partially inserted inside of the DNA duplex, which induces the buckle of the base pairs in vicinity (*27, 160, 181*). Third, it has also been reported that in hOGG1/DNA complex, a structurally similar residue, Y203, shows similar structural role in the structure distortion(*161*). For Fpg, the wedge is present in nearly identical intercalating locations with intact A:T and G:C base pairs (pdb id: 2F5P and 2F5O) as well as in the complex with everted 8-oxoG. Even though Fpg and hOGG1 adopt entirely different folds, by comparing the DNA in the enzyme/DNA complexes we observed that the conserved Y203 is positioned to play the same mechanistic role in hOGG1 as F114 in Fpg, with the aromatic rings occupying nearly identical locations with extrahelical 8-oxoG DNA (Figure 6-2). In both cases, the widowed cytosine forms two hydrogen bonds with a arginine in the vicinity (R112 in Fpg, R204 in hOGG1).

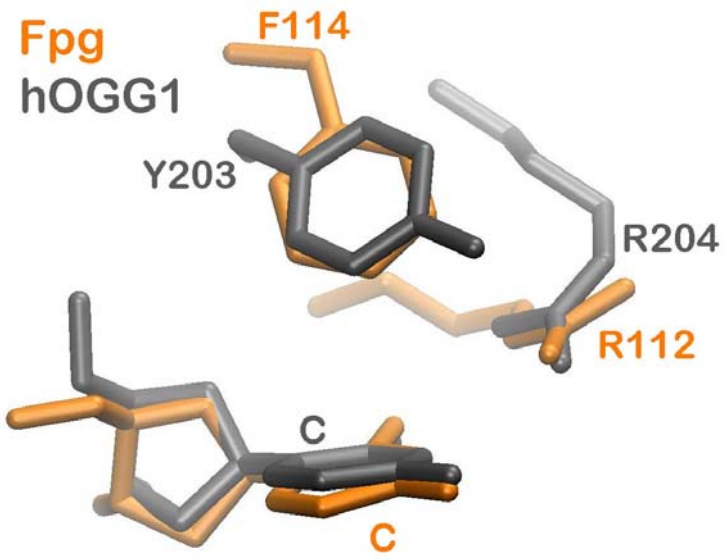

Figure 6-2. The overlap of the widowed cytosine and two neighboring key residues from x-ray structures of Fpg/DNA complex (PDB code: 1R2Y) and hOGG1/DNA complex (PDB code: 1YQR). The Fpg/DNA complex is shown in orange, and the hOGG1/DNA complex is shown in black.

In the available x-ray structures of Fpg/DNA complex, the aromatic ring of Phe114 is always inserted inside of the duplex. However, x-ray structure cannot exhibit the high energy transition state, such as how Phe114 moves from one site to the other site. One thing we can be certain about is that due to the steric effect, the aromatic ring of Phe114 has to be pulled out of the duplex for further sliding to occur, or the base pair has to break for Phe114 to go through. . The base pair breaking may need about 15 kcal/mol energy for breaking the three hydrogen bonds. the single molecule experiment have shown that the energy barrier for the Fpg/DNA translocation is about 2 kcal/mol(*176*), which is singnificantly below the necessary energy for breaking three hydrogen

bonds. It seems that the extrusion of the wedge residue Phe114 will be more feasible in this case. In this study, we measured the free energy profiles of pulling Phe114 from the original position with 8OG:C and G:C base pair present. The results show different energy barriers for the two systems. The energy barrier for pulling Phe114 out in OG:C system is higher than that in G:C system.

6.2 Methods

6.2.1 System preparation

All initial structures were built using the Leap module of Amber (version 9) (*128*), based on the crystal structure of the *B. st.* Fpg crosslinked with DNA (2F5O.pdb)(*160*). The sequence of the DNA duplex was 5'-AGGTAGACCTGGACGC-3', 5'-TGCGTCCG*GATCTACC-3' (where G* is at the interrogating site). All water molecules in the crystal structure were retained. The DNA and protein mutants were generated by manual editing of the pdb file, with the new side chain built using Leap. The TIP3P model(*129*) was used to explicitly represent water molecules. Following previous studies(*78, 125, 126*), the N-terminal proline was modeled as neutral to mimic the stage directly before the reaction. The parameters for neutral N-terminal proline were obtained from Perlow-Poehnelt *et al.* (*126*). Force field parameters for 8OG were obtained from Miller *et al*. The remaining protein and nucleic acid parameters employed Amber ff99 (*100, 132*), with modified protein backbone parameters to reduce the alpha-helical bias of those force fields(*133*). The initial coordinates were minimized and equilibrated following the procedures described in section 5.2.1.

6.2.2 Umbrella sampling

The umbrella sampling calculations were done at 330 K, the optimum temperature for *B. st.* bacterium. The initial structure was the structure after standard minimization and equilibration. The reaction coordinate was the distance between the mass center of the aromatic ring of Phe114 and the mass center of the two base pairs surrounding Phe114. The window size was 0.25 Å. The force constant was 60 kcal/mol×Å^2. For each window, a short simulation (5ps) was performed to generate the starting structures. The production runs were 2000 ps. The error bars were calculated using the second half of the data.

6.2.3 Electrostatic energy calculation

$$E = \sum \frac{q_i q_{O8}}{4\pi\varepsilon_0 r} \qquad 6.1$$

The Coulomb equation (6.1) is used in the electrostatic energy calculation. As shown in Figure 6-3, the electrostatic interactions between O8 atom and the atoms in the two adjacent phosphates and the oxygen atoms were calculated using equation 6.1. The distances between O8 and the other atoms were measured using ptraj. In this case we only need qualitative comparison of the electrostatic interaction of O8 at different positions, so we used vacuum environment for the calculation (dielectric constant $\varepsilon_r = 1$). The charges of the atoms were obtained from AMBER ff99. The charge of O8 atom of

8OG were obtained from Miller *et al*.(*130*). The values are shown in Table 6-1.

Atom	Charge (a.u.)
O3’	-0.5232
O4’	-0.3691
O5’	-0.4954
O1P	-0.7761
O2P	-0.7761
P	1.1659
O8	-0.5558

Table 6-1. The charges of the atom O8 of 8OG and the atoms in its neighboring phosphates.

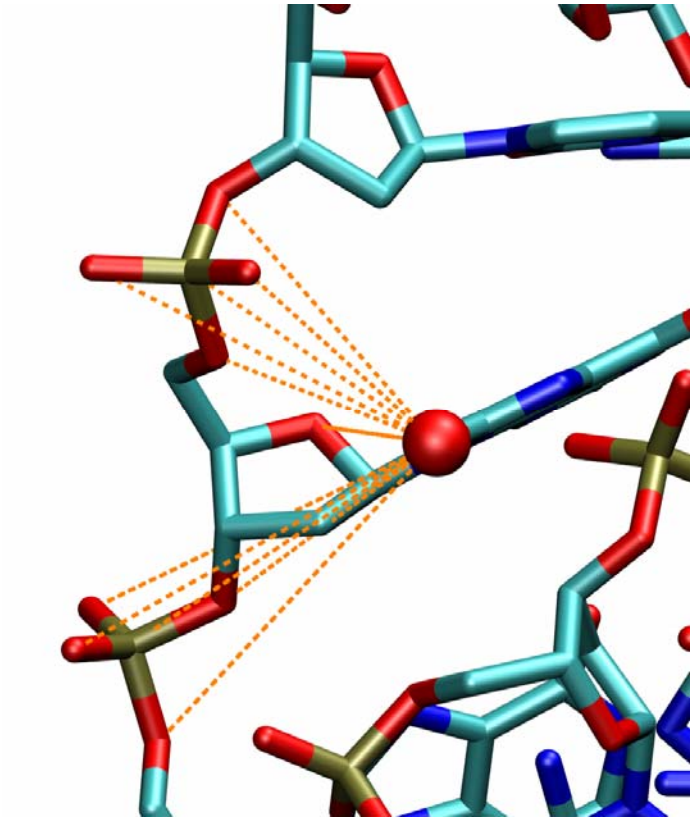

Figure 6-3. The atom O8 of 8OG and its neighboring charged atoms.

6.3 Results and discussion

6.3.1 The structural similarity between the two sampled ensembles

Before we calculated the potential of mean force from the sampled structure, we first examined the structural similarity between the structures sampled for the 8OG:C and G:C system. Four values, bending angle, buckle angle, chi angle and gamma angle were calculated for both systems. Bending angle represents the local backbone conformation of the DNA duplex. The buckle angle between 8OG:C or G:C base pair shows the

conformation of the base groups. The chi and gamma angles are the indicators of configuration of the nucleotide 8OG or G. To clarify the presentation, only the running averages over every 5 ps of the values are shown. The results are shown in Figure 6-4. We can see that all four values are similar between the two systems. In the two sets of the simulations, the two systems were sampling the similar conformations.

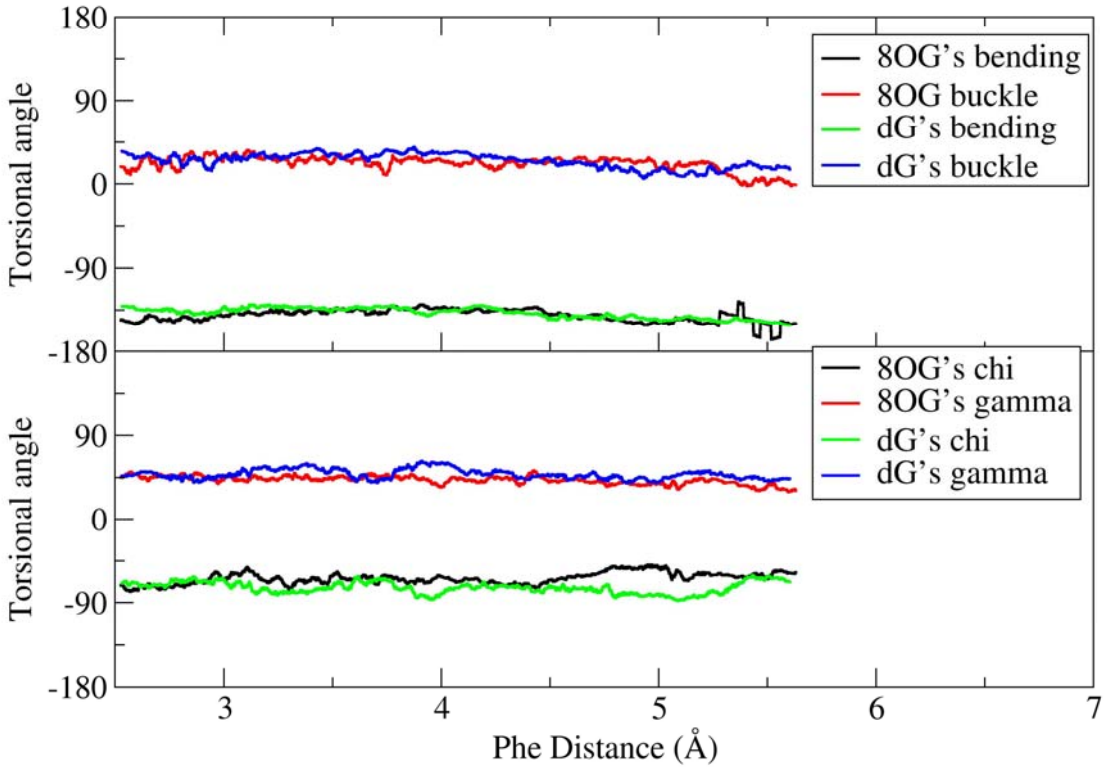

Figure 6-4. The structural characters of the two ensembles sampled in the two umbrella sampling simulations. The top panel shows the comparisons of bending angle and buckle angles of 8OG and dG systems. The bottom panel

shows the chi and gamma angles of the 8OG and dG in their systems. Bending angle represents the local backbone conformation of the DNA duplex. The buckle angle between 8OG:C or G:C base pair shows the conformation of the base groups. The chi and gamma angles are the indicators of configuration of the nucleotide 8OG or G. To clarify the presentation, only the running averages of the values are shown.

6.3.2 PMF simulation results

Figure 6-5 shows the free energy profiles for both 8OG:C and G:C systems. On the top of the figure, we show three snapshots of the wedge residue Phe114 and two base pairs in the vicinity. The base pair colored by the atom type are 8GO:C base pair, which is on the top of the wedge residue in the current view. The snapshot in the middle shows the structure at the free energy minimum. In the conformation, the wedge residue Phe114 is partially inserted into the duplex, and the interrogated base pair is buckled due to the wedge insertion. Deeper insertion of the wedge (shown in the snapfot on the left) will increase the free energy of the system, which is probably due to the steric interactions or the loss of the base stacking effects due to the more buckled base pair. The snapshot on the right shows the conformation at 4.5 Å. We visually examined the structures sampled in our simulation and found that in this conformation the aromatic ring of the wedge residue is completely out of the duplex and free to rotate. Once the wedge rotates its aromatic, there will be enough space for the translocation to proceed. Therefore, 4.5 Å

was chosen as the energy barrier for the wedge extrusion.

We can see that the location of the energy minimum of 8OG:dC is about the same position as that of dG:dC, which is between 3.5 and 3.75 Å. However, the heights of the energy barrier for the wedge residue Phe114 extrusion, which locates at 4.5 Å, are different between these two systems. The energy barrier for 8OG:C system is 3.5 kcal/mol, and that of G:C system is 2.3 kcal/mol. In a recently single molecular experimental study, it has been shown that for Fpg/DNA complex the translocation energy barrier for each base pair is about 2 kcal/mol(*176*), which is similar as the energy barrier shown in Figure 6-5 for the intact G:C base pair.

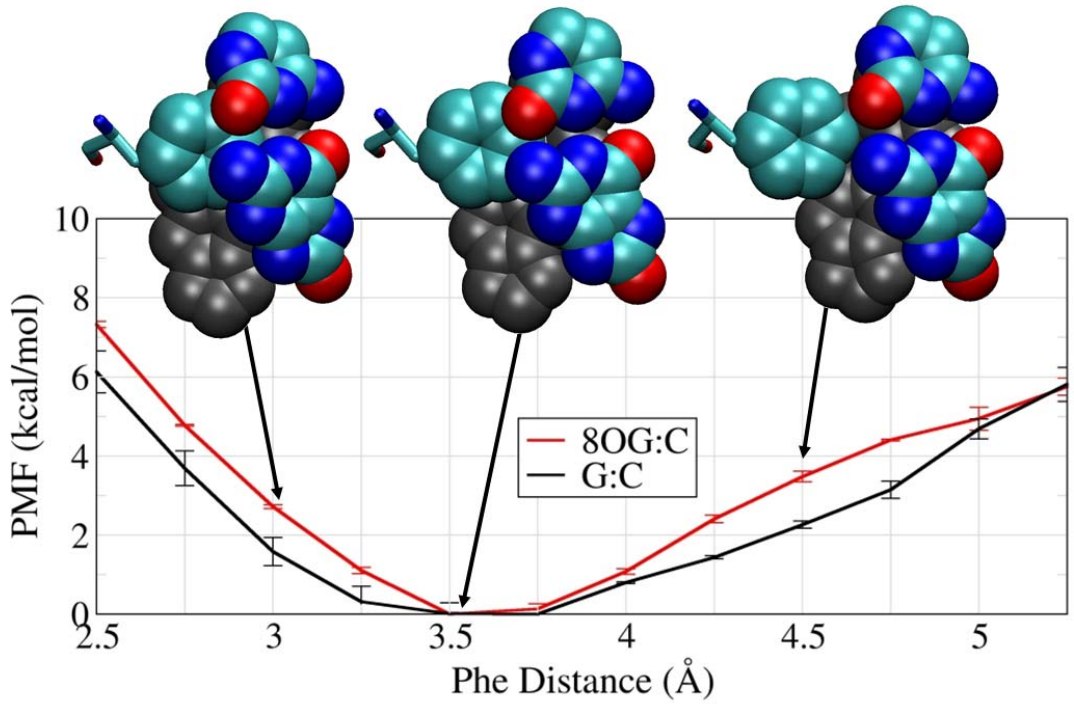

Figure 6-5. The free energy profiles for Phe114 insertion for Fpg/DNA containing 8OG:C base pair (in red) and Fpg/DNA containing G:C base pair (in black). Three snapshots representing the corresponding conformations are also shown on the top of the free energy profile.

The 1.2 kcal/mol free energy difference causes the rate of pulling Phe114 6 times slower in 8OG:C system than in G:C system according to the Arrhenius equation. Since DNA sliding cannot continue when the aromatic ring of Phe114 is trapped in DNA duplex, the slowing pulling rate suggests that 8OG:C base pair will have longer time staying in the interrogating site. Statistically this will create a larger window allowing further steps such as base pair breaking and base flipping to occur. By doing so, Phe114

kinetically discriminates 8OG from normal guanine.

6.3.3 The electrostatic interaction between O8 and neighboring phosphates

It is interesting to see that 8OG and normal guanine have different free energy profiles, in spite of their structural similarities (Figure 1-4). Previous simulation study has shown that the interaction between the atom O8 of 8OG and the phosphate atoms in the vicinity prefer 8OG's syn conformation over its anti conformation(*127*). One reasonable explanation for this difference is the electrostatic interaction between the atom O8 of 8OG with neighboring phosphates. Both O8 and the phosphate oxygens have negative charges, and they have repulsive interaction between each other. The insertion of Phe114 may change the distance between these atoms and therefore change the magnitude of the interactions.

To test this hypothesis, we measured the distances between O8 to all atoms in the two neighboring phosphates (Figure 6-3), and calculated the electrostatic energy using Coulomb equation (equation 6.1). The structures sampled in the window 3.5 Å (the free energy minimum) and 4.5 Å (the barrier) were analyzed. The histograms of the results are shown in Figure 6-6.

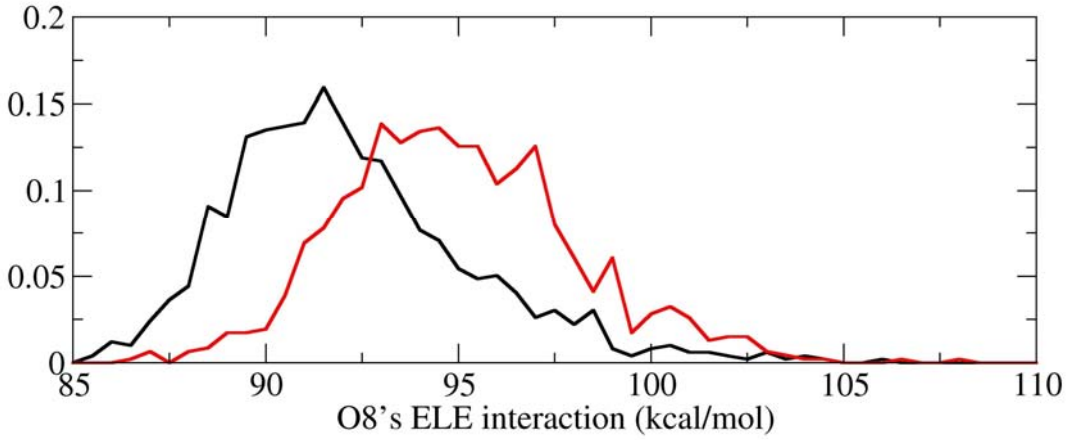

Figure 6-6. The histogram of the electrostatic interaction between O8 and neighboring other P and O atoms in the region of 3.25 ~ 3.75 Å (black) and 4.25 ~ 4.75 Å (red) of Phe114 distance.

In Figure 6-6 we can see that the populations of the electrostatic energy for the two windows have different average values. The average value for the structures in window 3.5 Å (shown in black) is about 4 kcal/mol lower than the one for window 4.5 Å (shown in red). It has been shown in Figure 6-4 OG:C and G:C systems sampled similar structures in the umbrella simulation. Therefore, we can assume that the energy difference between the two systems is caused solely by the interaction involving O8 (there is no significantly interaction involving N7). That fact that the electrostatic interactions between O8 and neighboring atoms stabilizes inserted structure and destabilizes the structure at the energy barrier can explain the energy difference of 8OG:C and G:C systems at the energy barrier. Although, the absolute value may not be

exact, due to the absence of the dielectric effect of the solvent in this calculation.

6.4 Conclusion

Crystalline crosslinked complexes of Fpg interrogating intrahelical A:T or G:C pairs reveal that the aromatic ring of F114 adopts an intercalating position, resulting in buckling of the base pair (*160*). As it is highly unlikely that each base pair is broken during sliding (*176*), F114 must be sufficiently extrahelical to "clear" each base pair during this process. We performed 2 sets of US simulations using the equilibrated structures of Fpg bound to DNA with intrahelical G:C or 8-oxoG:C pairs. For each system, we calculated the free energy as a function of the extent of intercalation. Insertion is favorable for G:C with a minimum at 3.5Å and the wedge inducing buckling and loss of stacking as observed experimentally (2F5O). We visually estimate that a distance of 4.5Å is sufficient to allow sliding to occur without opening of the G:C pair. Fpg wedge removal requires 2.3 kcal/mol for G:C pairs, comparable in magnitude to the sliding activation energy of 2.0 kcal/mol for Fpg as determined by single-molecule experiments (*176*). Importantly, the calculated barrier for wedge removal is significantly higher for 8-oxoG:C (3.5 kcal/mol). This analysis suggests that the rate for sliding past 8-oxoG:C pairs will be slower, providing additional time for eversion of the lesion as compared to undamaged bases, which provides a kinetic approcach for the lesion recognition. This effect may serve as one of the initial steps for lesion discrimation.

Bibliography

1. Collins, F. S., Lander, E. S., Rogers, J., Waterston, R. H., and Conso, I. H. G. S. (2004) Finishing the euchromatic sequence of the human genome, *Nature 431*, 931-945.

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

3. Wyatt, G. R. (1952) The Nucleic Acids of Some Insect Viruses, *J Gen Physiol 36*, 201-205.

4. Crick, F. (1974) Double Helix - a Personal View, *Nature 248*, 766-769.

5. Lodish, H. B., A.; Matsudaira, P.; Kaiser, C.A.; Krieger, M.; Scott, M.P.; Zipursky, S.L.; Darnell, J. (2004) *Molecular Cell Biology*, 5th ed., W.H. Freeman, New York.

6. E.C. Friedberg, G.C. Walker, W. Siede, R.D Wood, R. Shultz, and Ellenberger, T. (2006) *DNA Repair and Mutagenesis*, ASM Press, Washington, D.C.

7. Ohta, T., and Kimura, M. (1971) Functional Organization of Genetic Material as a Product of Molecular Evolution, *Nature 233*, 118-&.

8. Friedberg, E. C. (2003) DNA damage and repair, *Nature 421*, 436-440.

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

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

11. Bjelland, S., and Seeberg, E. (2003) *Mutagenicity, toxicity and repair of DNA base damage induced by oxidation*, Vol. 531.

12. Marnett, L. J. (2000) Oxyradicals and DNA damage, *Carcinogenesis 21*, 361-370.

13. Demple, B., and Harrison, L. (1994) Repair of Oxidative Damage to DNA -

Enzymology and Biology, *Annual Review of Biochemistry 63*, 915-948.

14. Ames, B. N., Gold, L. S., and Willett, W. C. (1995) The Causes and Prevention of Cancer, *Proceedings of the National Academy of Sciences of the United States of America 92*, 5258-5265.

15. Michaels, M. L., Tchou, J., Grollman, A. P., and Miller, J. H. (1992) A Repair System for 8-Oxo-7,8-Dihydrodeoxyguanine, *Biochemistry-Us 31*, 10964-10968.

16. Fowler, R. G., White, S. J., Koyama, C., Moore, S. C., Dunn, R. L., and Schaaper, R. M. (2003) Interactions among the Escherichia coli mutT, mutM, and mutY damage prevention pathways, *DNA Repair 2*, 159-173.

17. Chetsanga, C. J., and Lindahl, T. (1979) 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.

18. Cabrera, M., Nghiem, Y., and Miller, J. H. (1988) Mutm, a 2nd Mutator Locus in Escherichia-Coli That Generates G.C-]T.A Transversions, *Journal of Bacteriology 170*, 5405-5407.

19. Tchou, J., Kasai, H., Shibutani, S., Chung, M. H., Laval, J., Grollman, A. P., and Nishimura, S. (1991) 8-Oxoguanine (8-Hydroxyguanine) DNA Glycosylase and Its Substrate-Specificity, *Proceedings of the National Academy of Sciences of the United States of America 88*, 4690-4694.

20. Tchou, J., and Grollman, A. P. (1995) 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.

21. Zharkov, D. O., Rieger, R. A., Iden, C. R., and Grollman, A. P. (1997) 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.

22. Grollman, A. P., and Moriya, M. (1993) Mutagenesis by 8-Oxoguanine - an Enemy Within, *Trends in Genetics 9*, 246-249.

23. Bhagwat, M., and Gerlt, J. A. (1996) 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.

24. Tchou, J., Bodepudi, V., Shibutani, S., Antoshechkin, I., Miller, J., Grollman, A. P., and Johnson, F. (1994) Substrate-Specificity of Fpg Protein - Recognition and Cleavage of Oxidatively Damaged DNA, *Journal of Biological Chemistry 269*,

15318-15324.

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

26. Zharkov, D. O., Golan, G., Gilboa, R., Fernandes, A. S., Gerchman, S. E., Kycia, J. H., Rieger, R. A., Grollman, A. P., and Shoham, G. (2002) Structural analysis of an Escherichia coli endonuclease VIII covalent reaction intermediate, *Embo Journal 21*, 789-800.

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

28. Fromme, J. C., Bruner, S. D., Yang, W., Karplus, M., and Verdine, G. L. (2003) Product-assisted catalysis in base-excision DNA repair, *Nature Structural Biology 10*, 204-211.

29. Fromme, J. C., and Verdine, G. L. (2003) Structure of a trapped endonuclease III-DNA covalent intermediate, *Embo Journal 22*, 3461-3471.

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

31. van Gunsteren, W. F., Bakowies, D., Baron, R., Chandrasekhar, I., Christen, M., Daura, X., Gee, P., Geerke, D. P., Glattli, A., Hunenberger, P. H., Kastenholz, M. A., Ostenbrink, C., Schenk, M., Trzesniak, D., van der Vegt, N. F. A., and Yu, H. B. (2006) Biomolecular modeling: Goals, problems, perspectives, *Angewandte Chemie-International Edition 45*, 4064-4092.

32. Karplus, M., and McCammon, J. A. (2002) Molecular dynamics simulations of biomolecules, *Nature Structural Biology 9*, 646-652.

33. Jorgensen, W. L. (2004) The many roles of computation in drug discovery, *Science 303*, 1813-1818.

34. Simmerling, C., Elber, R. and Zhang, J. (1995) *MOIL-View - A Program for Visualization of Structure and Dynamics of Biomolecules and STO- A Program for Computing Stochastic Paths, in Modelling of Biomolecular Structure and Mechanisms*, , Kluwer, Netherlands

35. DeLano, W. L. (2002) The PyMOL Molecular Graphics System, DeLano

Scientific, San Carlos, CA, USA.

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

37. Plum, G. E., Grollman, A. P., Johnson, F., and Breslauer, K. J. (1995) Influence of the oxidatively damaged adduct 8-oxodeoxyguanosine on the conformation, energetics, and thermodynamic stability of a DNA duplex, *Biochemistry 34*, 16148-16160.

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

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

40. Beveridge, D. L., and Dicapua, F. M. (1989) Free-Energy Via Molecular Simulation - Applications to Chemical and Biomolecular Systems, *Annual Review of Biophysics and Biophysical Chemistry 18*, 431-492.

41. Zwanzig, R. W. (1954) High-Temperature Equation of State by a Perturbation Method .1. Nonpolar Gases, *Journal of Chemical Physics 22*, 1420-1426.

42. Alder, B. J., and Wainwright, T. E. (1957) Phase Transition for a Hard Sphere System, *Journal of Chemical Physics 27*, 1208-1209.

43. Stilling.Fh, and Rahman, A. (1974) Improved Simulation of Liquid Water by Molecular-Dynamics, *Journal of Chemical Physics 60*, 1545-1557.

44. Mccammon, J. A., Gelin, B. R., and Karplus, M. (1977) Dynamics of Folded Proteins, *Nature 267*, 585-590.

45. Okur, A., Strockbine, B., Hornak, V., and Simmerling, C. (2003) Using PC clusters to evaluate the transferability of molecular mechanics force fields for proteins, *Journal of Computational Chemistry 24*, 21-31.

46. Geney, R., Layten, M., Gomperts, R., Hornak, V., and Simmerling, C. (2006) Investigation of salt bridge stability in a generalized born solvent model, *Journal of Chemical Theory and Computation 2*, 115-127.

47. Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983) Optimization

by Simulated Annealing, *Science 220*, 671-680.

48. Hornak, V., and Simmerling, C. (2004) Development of softcore potential functions for overcoming steric barriers in molecular dynamics simulations, *Journal of Molecular Graphics & Modelling 22*, 405-413.

49. Freddolino, P. L., Arkhipov, A. S., Larson, S. B., McPherson, A., and Schulten, K. (2006) Molecular dynamics simulations of the complete satellite tobacco mosaic virus, *Structure 14*, 437-449.

50. Stern, H. A., Kaminski, G. A., Banks, J. L., Zhou, R. H., Berne, B. J., and Friesner, R. A. (1999) Fluctuating charge, polarizable dipole, and combined models: Parameterization from ab initio quantum chemistry, *Journal of Physical Chemistry B 103*, 4730-4737.

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

52. Darden, T., York, D., and Pedersen, L. (1993) Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems, *Journal of Chemical Physics 98*, 10089-10092.

53. Hockney, R. W., and Eastwood, J. W. (1988) *Computer simulation using particles*, Special student ed., A. Hilger, Bristol [England] ; Philadelphia.

54. Schlitter, J., Engels, M., Krüger, P., Jacoby, E., and Wollmer, A. (1993) Targeted molecular dynamics simulations of conformational change - application to the T↔R transition in insulin, *Molecular Simulation 19*, 291-309.

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

56. Wang, Y., Schulten, K., and Tajkhorshid, E. (2005) What makes an aquaporin a glycerol channel? A comparative study of AqpZ and GlpF, *Structure with Folding & Design 13*, 1107-1118.

57. Sanbonmatsu, K. Y., Joseph, S., and Tung, C. S. (2005) Simulating movement of tRNA into the ribosome during decoding, *Proceedings of the National Academy of Sciences U.S.A. 102*, 15854-15859.

58. Zhang, D. Q., Gullingsrud, J., and McCammon, J. A. (2006) Potentials of mean force for acetylcholine unbinding from the α7 nicotinic acetylcholine receptor

ligand-binding domain, *Journal of the American Chemical Society 128*, 3019-3026.

59. Yang, L. J., Beard, W. A., Wilson, S. H., Broyde, S., and Schlick, T. (2002) Polymerase b simulations suggest that Arg258 rotation is a slow step rather than large subdomain motions per se, *Journal of Molecular Biology 317*, 651-671.

60. Jensen, M. O., Park, S., Tajkhorshid, E., and Schulten, K. (2002) Energetics of glycerol conduction through aquaglyceroporin GlpF, *Proceedings of the National Academy of Sciences U.S.A. 99*, 6731-6736.

61. Amaro, R., Tajkhorshid, E., and Luthey-Schulten, Z. (2003) Developing an energy landscape for the novel function of a $(\beta/\alpha)_8$ barrel: Ammonia conduction through HisF, *Proceedings of the National Academy of Sciences U.S.A. 100*, 7599-7604.

62. Sugita, Y., and Okamoto, Y. (2000) An analysis on protein folding problem by replica-exchange method, *Prog Theor Phys Supp*, 402-403.

63. Sugita, Y., and Okamoto, Y. (1999) Replica-exchange molecular dynamics method for protein folding, *Chemical Physics Letters 314*, 141-151.

64. Earl, D. J., and Deem, M. W. (2005) Parallel tempering: Theory, applications, and new perspectives, *Physical Chemistry Chemical Physics 7*, 3910-3916.

65. Wickstrom, L., Okur, A., Song, K., Hornak, V., Raleigh, D. P., and Simmerling, C. L. (2006) The unfolded state of the villin headpiece helical subdomain: Computational studies of the role of locally stabilized structure, *Journal of Molecular Biology 360*, 1094-1107.

66. Roe, D. R., Hornak, V., and Simmerling, C. (2005) Folding cooperativity in a three-stranded beta-sheet model, *Journal of Molecular Biology 352*, 370-381.

67. Garcia, A. E., and Sanbonmatsu, K. Y. (2002) alpha-Helical stabilization by side chain shielding of backbone hydrogen bonds, *Proceedings of the National Academy of Sciences of the United States of America 99*, 2782-2787.

68. Zhou, R. H., Berne, B. J., and Germain, R. (2001) The free energy landscape for beta hairpin folding in explicit water, *Proceedings of the National Academy of Sciences of the United States of America 98*, 14931-14936.

69. Garcia, A. E., and Onuchic, J. N. (2003) Folding a protein in a computer: An atomic description of the folding/unfolding of protein A, *Proceedings of the National Academy of Sciences of the United States of America 100*, 13898-13903.

70. Zhou, R. H. (2003) Trp-cage: Folding free energy landscape in

explicit water, *Proceedings of the National Academy of Sciences of the United States of America 100*, 13280-13285.

71. Cecchini, M., Rao, F., Seeber, M., and Caflisch, A. (2004) Replica exchange molecular dynamics simulations of amyloid peptide aggregation, *Journal of Chemical Physics 121*, 10748-10756.

72. Tsai, H. H., Reches, M., Tsai, C. J., Gunasekaran, K., Gazit, E., and Nussinov, R. (2005) Energy landscape of amyloidogenic peptide oligomerization by parallel-tempering molecular dynamics simulation: Significant role of Asn ladder, *Proceedings of the National Academy of Sciences of the United States of America 102*, 8174-8179.

73. Verkhivker, G. M., Rejto, P. A., Bouzida, D., Arthurs, S., Colson, A. B., Freer, S. T., Gehlhaar, D. K., Larson, V., Luty, B. A., Marrone, T., and Rose, P. W. (2001) Parallel simulated tempering dynamics of ligand-protein binding with ensembles of protein conformations, *Chemical Physics Letters 337*, 181-189.

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

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

76. Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., and Kollman, P. A. (1995) Multidimensional Free-Energy Calculations Using the Weighted Histogram Analysis Method, *Journal of Computational Chemistry 16*, 1339-1350.

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

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

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

80. Priyakumar, U. D., and MacKerell, A. D. (2006) 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.

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

82. Huang, N., Banavali, N. K., and MacKerell, A. D. (2003) 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.

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

84. Rafi, S. B., Cui, G. L., Song, K., Cheng, X. L., Tonge, P. J., and Simmerling, C. (2006) Insight through molecular mechanics Poisson-Boltzmann surface area calculations into the binding affinity of triclosan and three analogues for FabI, the E-coli enoyl reductase, *Journal of Medicinal Chemistry 49*, 4574-4580.

85. Doig, A. J. (2002) Recent advances in helix-coil theory, *Biophysical Chemistry 101*, 281-293.

86. Barlow, D. J., and Thornton, J. M. (1988) Helix Geometry in Proteins, *Journal of Molecular Biology 201*, 601-619.

87. Schellman, J. A. (1959) The Factors Affecting the Stability of Hydrogen-Bonded Polypeptide Structures in Solution, *Journal of Physical Chemistry 62*, 1485-1494.

88. Zimm, B. H., and Bragg, J. K. (1959) Theory of the Phase Transition between Helix and Random Coil in Polypeptide Chains, *Journal of Chemical Physics 31*, 526-535.

89. Doty, P., and Yang, J. T. (1956) Polypeptides .7. Poly-Gamma-Benzyl-L-Glutamate - the Helix-Coil Transition in Solution, *Journal of the American Chemical Society 78*, 498-500.

90. Pan, K. M., Baldwin, M., Nguyen, J., Gasset, M., Serban, A., Groth, D., Mehlhorn, I., Huang, Z. W., Fletterick, R. J., Cohen, F. E., and Prusiner, S. B. (1993) Conversion of Alpha-Helices into Beta-Sheets Features in the Formation of the Scrapie Prion Proteins, *Proceedings of the National Academy of Sciences of the United States of America 90*, 10962-10966.

91. Jayawickrama, D., Zink, S., Vandervelde, D., Effiong, R. I., and Larive, C. K. (1995) Conformational-Analysis of the Beta-Amyloid Peptide Fragment, Beta(12-28), *Journal of Biomolecular Structure & Dynamics 13*, 229-244.

92. Marqusee, S., and Baldwin, R. L. (1987) Helix Stabilization by Glu- ... Lys+ Salt Bridges in Short Peptides of Denovo Design, *Proceedings of the National*

Academy of Sciences of the United States of America 84, 8898-8902.

93. Zhang, W., Lei, H. X., Chowdhury, S., and Duan, Y. (2004) Fs-21 peptides can form both single helix and helix-turn-helix, *Journal of Physical Chemistry B 108*, 7479-7489.

94. Nymeyer, H., and Garcia, A. E. (2003) Simulation of the folding equilibrium of alpha-helical peptides: A comparison of the generalized born approximation with explicit solvent, *Proceedings of the National Academy of Sciences of the United States of America 100*, 13934-13939.

95. Jas, G. S., and Kuczera, K. (2004) Equilibrium structure and folding of a helix-forming peptide: Circular dichroism measurements and replica-exchange molecular dynamics simulations, *Biophysical Journal 87*, 3786-3798.

96. Scheraga, H. A., Vile, J. A., and Ripoll, D. R. (2002) Helix-coil transitions re-visited, *Biophysical Chemistry 101*, 255-265.

97. Lin, J. C., Barua, B., and Andersen, N. H. (2004) The helical alanine controversy: An (Ala)(6) insertion dramatically increases helicity, *Journal of the American Chemical Society 126*, 13679-13684.

98. Schneider, J. P., and DeGrado, W. F. (1998) The design of efficient alpha-helical C-capping auxiliaries, *Journal of the American Chemical Society 120*, 2764-2767.

99. Huang, C. Y., Klemke, J. W., Getahun, Z., DeGrado, W. F., and Gai, F. (2001) Temperature-dependent helix-coil transition of an alanine based peptide, *Journal of the American Chemical Society 123*, 9235-9238.

100. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995) A 2Nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules, *Journal of the American Chemical Society 117*, 5179-5197.

101. Zaman, M. H., Shen, M. Y., Berry, R. S., Freed, K. F., and Sosnick, T. R. (2003) Investigations into sequence and conformational dependence of backbone entropy, inter-basin dynamics and the flory isolated-pair hypothesis for peptides, *Journal of Molecular Biology 331*, 693-711.

102. Hawkins, G. D., Cramer, C. J., and Truhlar, D. G. (1995) Pairwise Solute Descreening of Solute Charges from a Dielectric Medium, *Chemical Physics Letters 246*, 122-129.

103. Hawkins, G. D., Cramer, C. J., and Truhlar, D. G. (1996) Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic

charges from a dielectric medium, *Journal of Physical Chemistry 100*, 19824-19839.

104. Tsui, V., and Case, D. A. (2000) Theory and applications of the generalized Born solvation model in macromolecular Simulations, *Biopolymers 56*, 275-291.

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

106. Roccatano, D., Amadei, A., Di Nola, A., and Berendsen, H. J. C. (1999) A molecular dynamics study of the 41-56 beta-hairpin from B1 domain of protein G, *Protein Science 8*, 2130-2143.

107. Hong, Q., and Schellman, J. A. (1992) Helix-Coil Theories - a Comparative-Study for Finite Length Polypeptides, *Journal of Physical Chemistry 96*, 3987-3994.

108. Lifson, S. (1961) Theory of Helix-Coil Transition in Polypeptides, *Journal of Chemical Physics 34*, 1963-&.

109. Sorin, E. J., and Pande, V. S. (2005) Exploring the helix-coil transition via all-atom equilibrium ensemble simulations, *Biophysical Journal 88*, 2472-2493.

110. Kabsch, W., and Sander, C. (1983) Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features, *Biopolymers 22*, 2577-2637.

111. Werner, J. H., Dyer, R. B., Fesinmeyer, R. M., and Andersen, N. H. (2002) Dynamics of the primary processes of protein folding: Helix nucleation, *Journal of Physical Chemistry B 106*, 487-494.

112. Scholtz, J. M., and Baldwin, R. L. (1992) The Mechanism of Alpha-Helix Formation by Peptides, *Annu Rev Bioph Biom 21*, 95-118.

113. Groebke, K., Renold, P., Tsang, K. Y., Allen, T. J., McClure, K. F., and Kemp, D. S. (1996) Template-nucleated alanine-lysine helices are stabilized by position-dependent interactions between the lysine side chain and the helix barrel, *Proc Natl Acad Sci U S A 93*, 4025-4029.

114. E.C. Frieberg, G.C. Walker, W. Siede, R.D Wood, R. Shultz, and Ellenberger, T. (2006) *DNA Repair and Mutagenesis*, ASM Press, Washington, D.C.

115. Halliwell, B., and Gutteridge, J. M. C. (1999) *Free radicals in biology and medicine*, 3rd ed., Oxford University Press, Oxford.

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

117. Tchou, J., and Grollman, A. P. (1993) Repair of DNA Containing the Oxidatively-Damaged Base, 8-Oxoguanine, *Mutation Research 299*, 277-287.

118. Michaels, M. L., and Miller, J. H. (1992) The Go System Protects Organisms from the Mutagenic Effect of the Spontaneous Lesion 8-Hydroxyguanine (7,8-Dihydro-8-Oxoguanine), *Journal of Bacteriology 174*, 6321-6325.

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

120. Sugahara, M., Mikawa, T., Kumasaka, T., Yamamoto, M., Kato, R., Fukuyama, K., Inoue, Y., and Kuramitsu, S. (2000) Crystal structure of a repair enzyme of oxidatively damaged DNA, MutM (Fpg), from an extreme thermophile, Thermus thermophilus HB8, *Embo Journal 19*, 3857-3869.

121. Fromme, J. C., and Verdine, G. L. (2003) DNA lesion recognition by the bacterial repair enzyme MutM, *Journal of Biological Chemistry 278*, 51543-51548.

122. Serre, L., de Jesus, K. P., Boiteux, S., Zelwer, C., and Castaing, B. (2002) Crystal structure of the Lactococcus lactis formamidopyrimidine-DNA glycosylase bound to an abasic site analogue-containing DNA, *Embo Journal 21*, 2854-2865.

123. Francis, A. W., Helquist, S. A., Kool, E. T., and David, S. S. (2003) Probing the requirements for recognition and catalysis in fpg and MutY with nonpolar adenine isosteres, *Journal of the American Chemical Society 125*, 16235-16242.

124. Lavrukhin, O. V., and Lloyd, R. S. (2000) Involvement of phylogenetically conserved acidic amino acid residues in catalysis by an oxidative DNA damage enzyme formamidopyrimidine glycosylase, *Biochemistry 39*, 15266-15271.

125. Zaika, E. I., Perlow, R. A., Matz, E., Broyde, S., Gilboa, R., Grollman, A. P., and Zharkov, D. O. (2004) Substrate discrimination by formamidopyrimidine-DNA glycosylase - A mutational analysis, *Journal of Biological Chemistry 279*, 4849-4861.

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

127. Cheng, X. L., Kelso, C., Hornak, V., de los Santos, C., Grollman, A. P., and Simmerling, C. (2005) Dynamic behavior of DNA base pairs containing 8-

oxoguanine, *Journal of the American Chemical Society 127*, 13906-13918.

128. Case, D. A. e. a. (2004) *AMBER 8*, University of California, San Francisco.

129. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of Simple Potential Functions for Simulating Liquid Water, *Journal of Chemical Physics 79*, 926-935.

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

131. Stote, R. H., and Karplus, M. (1995) Zinc-Binding in Proteins and Solution - a Simple but Accurate Nonbonded Representation, *Proteins-Structure Function and Genetics 23*, 12-31.

132. Wang, J. M., Cieplak, P., and Kollman, P. A. (2000) 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.

133. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. and Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters, *Proteins: Structure, Function and Genetics*, In Press.

134. Fuxreiter, M., Warshel, A., and Osman, R. (1999) Role of active site residues in the glycosylase step of T4 Endonuclease V. Computer simulation studies on ionization states, *Biochemistry 38*, 9577-9589.

135. Forsyth, W. R., Antosiewiez, J. M., and Robertson, A. D. (2002) Empirical relationships between protein structure and carboxyl pK(a) values in proteins, *Proteins-Structure Function and Genetics 48*, 388-403.

136. Li, H., Robertson, A. D., and Jensen, J. H. (2005) Very fast empirical prediction and rationalization of protein pK(a) values, *Proteins-Structure Function and Bioinformatics 61*, 704-721.

137. Gordon, J. C., Myers, J. B., Folta, T., Shoja, V., Heath, L. S., and Onufriev, A. (2005) H++: a server for estimating pK(a)s and adding missing hydrogens to macromolecules, *Nucleic Acids Research 33*, W368-W371.

138. Cheatham, T. E., Miller, J. L., Fox, T., Darden, T. A., and Kollman, P. A. (1995) 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.

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

140. Massova, I., and Kollman, P. A. (2000) Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding, *Perspectives in Drug Discovery and Design 18*, 113-135.

141. Feig, M., Onufriev, A., Lee, M. S., Im, W., Case, D. A., and Brooks, C. L. (2004) Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures, *Journal of Computational Chemistry 25*, 265-284.

142. Onufriev, A., Bashford, D., and Case, D. A. (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model, *Proteins-Structure Function and Bioinformatics 55*, 383-394.

143. Sitkoff, D., Sharp, K. A., and Honig, B. (1994) Accurate Calculation of Hydration Free-Energies Using Macroscopic Solvent Models, *Journal of Physical Chemistry 98*, 1978-1988.

144. Amara, P., Serre, L., Castaing, B., and Thomas, A. (2004) Insights into the DNA repair process by the formamidopyrimidine-DNA glycosylase investigated by molecular dynamics, *Protein Science 13*, 2009-2021.

145. Friedberg, E. C., Walker, G. C., Siede, W., Wood, R. D., Shultz, R., and Ellenberger, T. (2006) *DNA Repair and Mutagenesis*, ASM Press, Washington, D.C.

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

147. Lindahl, T. (1987) Regulation and Deficiencies in DNA-Repair, *British Journal of Cancer 56*, 91-95.

148. Delaney, M. O., and Greenberg, M. M. (2002) Synthesis of oligonucleotides and thermal stability of duplexes containing the beta-C-nucleoside analogue of Fapy center dot dG, *Chemical Research in Toxicology 15*, 1460-1465.

149. Haraguchi, K., Delaney, M. O., Wiederholt, C. J., Sambandam, A., Hantosi, Z., and Greenberg, M. M. (2002) Synthesis and characterization of oligodeoxynucleotides containing formamidopyrimidine lesions and nonhydrolyzable analogues, *Journal of the American Chemical*

Society 124, 3263-3269.

150. Wiederholt, C. J., Delaney, M. O., Pope, M. A., David, S. S., and Greenberg, M. M. (2003) Repair of DNA containing Fapy center dot dG and its beta-C-nucteoside analogue by formamidopyrimidine DNA glycosylase and MutY, *Biochemistry 42*, 9755-9760.

151. Patro, J. N., Haraguchi, K., Delaney, M. O., and Greenberg, M. M. (2004) Probing the configurations of formamidopyrimidine lesions Fapy center dot dA and Fapy center dot dG in DNA using endonuclease IV, *Biochemistry 43*, 13397-13403.

152. Coste, F., Ober, M., Carell, T., Boiteux, S., Zelwer, C., and Castaing, B. (2004) Structural basis for the recognition of the FapydG lesion (2,6-diamino-4-hydroxy-5-formamidopyrimidine) by formamidopyrimidine-DNA glycosylase, *Journal of Biological Chemistry 279*, 44074-44083.

153. Ober, M., Linne, U., Gierlich, J., and Carell, T. (2003) The two main DNA lesions 8-oxo-7,8-dihydroguanine and 2,6-diamino-5-formamido-4-hydroxypyrimidine exhibit strongly different pairing properties, *Angewandte Chemie-International Edition 42*, 4947-4951.

154. Ober, M., Muller, H., Pieck, C., Gierlich, J., and Carell, T. (2005) Base pairing and replicative processing of the formamidopyrimidine-dG DNA lesion, *Journal of the American Chemical Society 127*, 18143-18149.

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

156. Gaussian 03, R. C., Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A.; . Gaussian, Inc.,

Wallingford CT, 2004.

157. Cornell, W. D., Cieplak, P., Bayly, C. I., and Kollman, P. A. (1993) Application of Resp Charges to Calculate Conformational Energies, Hydrogen-Bond Energies, and Free-Energies of Solvation, *Journal of the American Chemical Society 115*, 9620-9631.

158. Berg, O. G., Winter, R. B., and Vonhippel, P. H. (1981) Diffusion-Driven Mechanisms of Protein Translocation on Nucleic-Acids .1. Models and Theory, *Biochemistry 20*, 6929-6948.

159. Gowers, D. M., Wilson, G. G., and Halford, S. E. (2005) Measurement of the contributions of 1D and 3D pathways to the translocation of a protein along DNA, *Proceedings of the National Academy of Sciences of the United States of America 102*, 15883-15888.

160. Banerjee, A., Santos, W. L., and Verdine, G. L. (2006) Structure of a DNA glycosylase searching for lesions, *Science 311*, 1153-1157.

161. Banerjee, A., Yang, W., Karplus, M., and Verdine, G. L. (2005) Structure of a repair enzyme interrogating undamaged DNA elucidates recognition of damaged DNA, *Nature 434*, 612-618.

162. Banerjee, A., and Verdline, G. L. (2006) A nucleobase lesion remodels the interaction of its normal neighbor in a DNA glycosylase complex, *Proceedings of the National Academy of Sciences of the United States of America 103*, 15020-15025.

163. Bredenberg, J., and Nilsson, L. (2001) Modeling zinc sulfhydryl bonds in zinc fingers, *International Journal of Quantum Chemistry 83*, 230-244.

164. Doruker, P., Atilgan, A. R., and Bahar, I. (2000) Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to alpha-amylase inhibitor, *Proteins-Structure Function and Genetics 40*, 512-524.

165. Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model, *Biophysical Journal 80*, 505-515.

166. Xu, C. Y., Tobi, D., and Bahar, I. (2003) Allosteric changes in protein structure computed by a simple mechanical model: Hemoglobin T <-> R2 transition, *Journal of Molecular Biology 333*, 153-168.

167. Wang, Y. M., Rader, A. J., Bahar, I., and Jernigan, R. L. (2004) Global ribosome motions revealed with elastic network model, *Journal of Structural Biology 147*,

302-314.

168. Keskin, O., Durell, S. R., Bahar, I., Jernigan, R. L., and Covell, D. G. (2002) Relating molecular flexibility to function: A case study of tubulin, *Biophysical Journal 83*, 663-680.

169. Kuznetsov, N. A., Koval, V. V., Zharkov, D. O., Vorobjev, Y. N., Nevinsky, G. A., Douglas, K. T., and Fedorova, O. S. (2007) Pre-steady-state kinetic study of substrate specificity of Escherichia coli formamidopyrimidine-DNA glycosylase, *Biochemistry 46*, 424-435.

170. Allan, B. W., Reich, N. O., and Beechem, J. M. (1999) Measurement of the absolute temporal coupling between DNA binding and base flipping, *Biochemistry 38*, 5308-5314.

171. Drohat, A. C., Jagadeesh, J., Ferguson, E., and Stivers, J. T. (1999) Role of electrophilic and general base catalysis in the mechanism of Escherichia coli uracil DNA glycosylase, *Biochemistry 38*, 11866-11875.

172. Bahar, I., and Rader, A. J. (2005) Coarse-grained normal mode analysis in structural biology, *Current Opinion in Structural Biology 15*, 586-592.

173. Ma, J. P. (2005) Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes, *Structure 13*, 373-380.

174. Ma, J. P. (2004) New advances in normal mode analysis of supermolecular complexes and applications to structural refinement, *Current Protein & Peptide Science 5*, 119-123.

175. Eyal, E., Yang, L. W., and Bahar, I. (2006) Anisotropic network model: systematic evaluation and a new web interface, *Bioinformatics 22*, 2619-2627.

176. Blainey, P. C., van Oijent, A. M., Banerjee, A., Verdine, G. L., and Xie, X. S. (2006) 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.

177. Burrows, C. J., and Muller, J. G. (1998) Oxidative nucleobase modifications leading to strand scission, *Chemical Reviews 98*, 1109-1151.

178. Kouchakdjian, M., Bodepudi, V., Shibutani, S., Eisenberg, M., Johnson, F., Grollman, A. P., and Patel, D. J. (1991) Nmr Structural Studies of the Ionizing-Radiation Adduct 7-Hydro-8-Oxodeoxyguanosine (8-Oxo-7H-Dg) Opposite Deoxyadenosine in a DNA Duplex - 8-Oxo-7H-Dg(Syn).Da(Anti) Alignment at Lesion Site, *Biochemistry 30*, 1403-1412.

179. Moriya, M., Ou, C., Bodepudi, V., Johnson, F., Takeshita, M., and Grollman, A. P. (1991) Site-Specific Mutagenesis Using a Gapped Duplex Vector - a Study of Translesion Synthesis Past 8-Oxodeoxyguanosine in Escherichia-Coli, *Mutation Research 254*, 281-288.

180. Moriya, M. (1993) Single-Stranded Shuttle Phagemid for Mutagenesis Studies in Mammalian-Cells - 8-Oxoguanine in DNA Induces Targeted G.C to T.A Transversions in Simian Kidney-Cells, *Proceedings of the National Academy of Sciences of the United States of America 90*, 1122-1126.

181. Bruner, S. D., Norman, D. P. G., and Verdine, G. L. (2000) Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA, *Nature 403*, 859-866.

Appendices

Appendix A – Principle component analysis

Principle component analysis is used to reduce to the dimentions of the conformational space. It can be done using ptraj in Amber. The following is the input file for carrying principle component analysis:

```
#!/bin/tcsh

set ptraj = "/mnt/raidb/kensong/amber8/exe/ptraj"

set top = "pept.mod2.bugfix.top"

set currDir = $PWD


EignVectors:

$ptraj $top << EOF

trajin ../RunREM/290.x  5000 40000

reference ./helix.crd

rms reference :5-20@CA

matrix covar name covarmax :5-20@CA

analyze matrix covarmax out covecs.dat vecs 48

EOF


grep -A 1 '[****]' covecs.dat | awk ' $2 > 0{print $2}' > eignvalue.dat


Projection:
```

```
$ptraj ../$top << EOF

trajin ../../GBREMTraj/GBREM.x

reference ../helix.crd

rms reference :5-20@CA

projection modes ../covecs.dat out pca.dat beg 1 end 2 :5-20@CA

EOF


FreeEne:

awk '{print $1+100, $2+100}' pca.dat > temp

2dhist temp 70 130 0.5 70 130 0.5 > temp2

awk '{print $1-100, $2-100, $3*0.00024}' temp2 > FreeEng$tem:r.dat


rm temp temp2

echo "----------------------------------------\n"

echo  "analysis finished!\n"

echo "----------------------------------------\n"
```

Appendix B – Radius of gyration analysis

```
#!/bin/tcsh

set top = ../pept.mod2.bugfix.top
set carnal = ~kensong/amber8.cmpi/exe/carnal
set i = 0
while ( $i < 12)
set ext=`printf "%3.3i" $i`

echo "trajectory "$ext

$carnal << EOF
    FILES_IN
    PARM p1 $top;
    STREAM s1 ../GBREMupto9.x.$ext;
    FILES_OUT
    TABLE tab1 gr$ext:rK.dat;
    DECLARE
    GROUP g1 ((RES 5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20) & (ATOM NAME CA));
    OUTPUT
    TABLE tab1 g1%radgyr;
    END
EOF

temp:
mv gr$ext:rK.dat temp
awk '{print $2}' temp  > GRDis$ext:r.dat
rm temp

@ i ++
end
```

Appendix C – relaxed potential energy scan

```
%nproc = 4
%chk=tor.checkpoint
# HF/6-31G* Opt=ModRedundant

FAG relaxed potential energy scan

0  1
O 25.055 50.042 21.398
C 26.168 50.841 21.042
C 26.658 51.796 22.140
O 25.590 52.683 22.456
C 25.565 52.848 23.872
C 25.924 51.469 24.403
N 24.319 53.337 24.419
C 23.798 54.576 24.186
C 22.443 54.900 24.358
C 21.895 56.101 23.855
O 20.739 56.361 23.644
N 22.888 57.045 23.506
C 24.205 56.706 23.424
N 25.042 57.654 23.044
N 21.548 54.001 24.775
C 21.342 52.849 24.158
O 22.035 52.329 23.287
N 24.676 55.479 23.713
C 27.088 51.159 23.477
O 28.235 51.833 23.964
H 25.908 51.428 20.161
H 26.993 50.193 20.762
H 27.503 52.362 21.744
H 26.347 53.542 24.158
H 25.108 50.760 24.247
H 26.224 51.512 25.449
H 23.707 52.606 24.803
H 22.584 57.976 23.287
H 24.728 58.610 22.930
H 26.044 57.482 23.081
H 20.810 54.283 25.405
H 20.413 52.409 24.404
H 27.254 50.082 23.417
H 24.816 49.478 20.659
H 28.508 51.441 24.796

* 9 15 * R
10 9 15 16 S 35 10.0
```

Appendix D – MP2 single point energy

#MP2/6-31G*

Single point calculation on Frame 0

0 1
O -3.4984 -1.7084 1.4888
C -3.6705 -0.3317 1.7697
C -3.0717 0.6225 0.7262
O -1.6700 0.3797 0.6659
C -1.2799 0.4175 -0.7050
C -2.4494 -0.2279 -1.4317
N -0.0227 -0.2277 -1.0116
C 1.1977 0.1930 -0.5706
C 2.3342 -0.6289 -0.5111
C 3.4955 -0.2437 0.1950
O 4.3894 -0.9427 0.5958
N 3.5098 1.1363 0.5022
C 2.4022 1.9164 0.3579
N 2.5134 3.1820 0.7182
N 2.3032 -1.9078 -0.8941
C 1.4608 -2.7923 -0.3860
O 0.4752 -2.5747 0.3144
N 1.2326 1.4636 -0.1297
C -3.5871 0.4854 -0.7208
O -3.7256 1.7763 -1.2878
H -3.2203 -0.1147 2.7384
H -4.7317 -0.1186 1.8558
H -3.2551 1.6458 1.0586
H -1.2065 1.4520 -1.0203
H -2.4895 -1.3026 -1.2411
H -2.4232 -0.0100 -2.4983
H -0.1136 -1.1815 -1.3836
H 4.3634 1.5204 0.8643
H 3.4107 3.5709 0.9812
H 1.7528 3.8279 0.5202
H 3.0962 -2.3059 -1.3775
H 1.7657 -3.7918 -0.5449
H -4.5117 -0.0883 -0.8051
H -3.9002 -2.2305 2.1870
H -4.0465 1.6979 -2.1886

Appendix E – Point charge fitting

In common MD simulations, the charges of the molecue are represented as the point charges at the centers of atoms. The standard way of calculating the point charges for new residues are: first, optimizing the structure and calculating the electrostatics potential surface (EPS); second, using RESP module of Amber to fit point charges to the EPS.
For optimizing the structure and calculating the EPS:

```
%chk=ESP.chk
# HF/6-31G* pop=(minimal, mk) opt iop(6/33=2)

optimize and ESP calculation

0  1
O      25.055  50.042  21.398
C      26.168  50.841  21.042
C      26.658  51.796  22.140
O      25.590  52.683  22.456
C      25.565  52.848  23.872
C      25.924  51.469  24.403
N      24.319  53.337  24.419
C      23.798  54.576  24.186
C      22.443  54.900  24.358
C      21.895  56.101  23.855
O      20.739  56.361  23.644
N      22.888  57.045  23.506
C      24.205  56.706  23.424
N      25.042  57.654  23.044
N      21.548  54.001  24.775
C      21.342  52.849  24.158
O      22.035  52.329  23.287
N      24.676  55.479  23.713
C      27.088  51.159  23.477
O      28.235  51.833  23.964
H      25.908  51.428  20.161
H      26.993  50.193  20.762
H      27.503  52.362  21.744
H      26.347  53.542  24.158
H      25.108  50.760  24.247
H      26.224  51.512  25.449
H      23.707  52.606  24.803
H      22.584  57.976  23.287
H      24.728  58.610  22.930
H      26.044  57.482  23.081
H      20.810  54.283  25.405
H      20.413  52.409  24.404
H      27.254  50.082  23.417
H      24.816  49.478  20.659
H      28.508  51.441  24.796
```

For RESP calculation:

--input file1 (option_1.in)

```
ESP for Fapy_DMP
 &cntrl
   inopt=0,
   ioutopt=0,
   iqopt=1,
   nmol=2,
   ihfree=1,
   irstrnt=1,
   qwt=0.0005
 &end
1.000000
ESP
    0   35
    8    0
    6    0
    6    0
    8    0
    6    0
    6    0
    7    0
    6    0
    6    0
    6    0
    8    0
    7    0
    6    0
    7    0
    7    0
    6    0
    8    0
    7    0
    6    0
    8    0
    1    0
    1    0
    1    0
    1    0
    1    0
    1    0
    1    0
    1    0
    1    0
    1   29
    1    0
    1    0
    1    0
    1    0
    1    0

1.000000
DMP
```

```
  -1   13
   6    0
   8    0
   1    0
   1    0
   1    0
  15    0
   8    0
   8    7
   8    0
   6    0
   1    0
   1    0
   1    0

   6   0.000
   1   20    1   35    2    1    2    3    2    4    2    5
   6   0.000
   1    1    1   34    2   10    2   11    2   12    2   13
```

--input file2 (option_2.in)

```
ESP for Fapy_DMP
 &cntrl
  inopt=0,
  ioutopt=0,
  iqopt=2,
  nmol=2,
  ihfree=1,
  irstrnt=1,
  qwt=0.0010
 &end
1.000000
ESP
   0   35
   8   -99
   6     0
   6   -99
   8   -99
   6   -99
   6     0
   7   -99
   6   -99
   6   -99
   6   -99
   8   -99
   7   -99
   6   -99
   7   -99
   7   -99
   6   -99
   8   -99
   7   -99
   6   -99
```

```
   8   -99
   1     0
   1    21
   1   -99
   1   -99
   1     0
   1    25
   1   -99
   1   -99
   1   -99
   1   -99
   1   -99
   1   -99
   1   -99
   1   -99

1.000000
DMP
  -1   13
   6   -99
   8   -99
   1   -99
   1   -99
   1   -99
  15   -99
   8   -99
   8   -99
   8   -99
   6   -99
   1   -99
   1   -99
   1   -99
```

And the run script for RESP:

```
#!/bin/csh -f
echo " "
echo "  Amber 8   RESP"
echo " "
echo RESP stage 1:
resp -O \
             -i option_1.in \
             -o FAG_1st.out \
             -p FAG_1st.pch \
             -e espot.dat \
             -t FAG_1st.chg              || goto error

echo RESP stage 2:
resp -O \
             -i option_2.in \
             -o FAG_2nd.out \
             -p FAG_2nd.pch \
```

```
        -e espot.dat \
        -t FAG_2nd.chg \
        -q FAG_1st.chg          || goto error

echo No errors detected
exit(0)

error:
echo Error: check .out and try again
exit(1)
```

Appendix F – Targeted MD simulation

Targeted MD simulation uses a guiding force which is proportional to the structure difference between the current structure and the targeting structure (RMSD value) to force the structure in the simulation to evolve to the targeted structure. In the applications, the tragetting RMSD value can be set to change gradually over the simulation, which can generate a more even force over the simulation. The following is an example of the input file for such type of targeted MD:

```
md.in
 &cntrl
       imin = 0, ntx = 5, nstlim = 250000,
       ntc = 2, ntf = 1, tol=0.0000001, ntt = 1, dt = 0.002,
       ntb = 2, ntp = 1, irest = 1,
       ntwx = 500, ntwe = 0, ntwr = 500, ntpr = 500,
       scee = 1.2, cut = 8.0, npscal = 1,
       ntr = 0, ibelly = 0, temp0 = 330.0,
       nscm = 5000, iwrap = 1,
       itgtmd = 1, tgtmdfrc = 5,
       tgtfitmask = ":1-273@CA", tgtrmsmask = "@4883-4998",
       nmropt = 1,
 &end
 &wt type = 'TGTRMSD', istep1 = 1, istep2 = 250000,
    value1 = 6.15, value2 = 0,
 &end
 &wt type = 'END'
 &end
```

Appendix G – Pseudo-dihedral angle calcualtion

Pseudo-dihedral angle is used to measure the position of flipping base. The following is the ptraj script for calculating the pseudo-dihedral angle:

```
#!/bin/tcsh

set ptraj = /mnt/raidb/programs/amber8.ifort/exe/ptraj
set top = ../../0GENE/nowat8OGIC3.parm7

$ptraj $top << EOF
trajin nowattraj1.x
trajin nowatmd2.x


dihedral
PHI291 :292@N9,C8,N7,C5,C6,O6,N1,C2,N2,N3,C4,:279@N1,C6,C5,C4,N4,N3,C2,O2 :292@C
1',C2',C3',C4' :291@C1',C2',C3',C4' :291@N1,C2,N3,C4,C5,C6  out PHI.dat

go
EOF
```

Appendix H – Howto run umbrella sampling simulation.

Umbrella sampling simulation is to calculate the free energy profile of one system along one (or multiple) reaction coordinate(s). In most cases, the number of reaction coordinate is one. It could also be two, which is called 2-D umbrella sampling. The higher dimension is rare. In our studies, we are using one reaction coordinate. 2-D umbrella sampling is using the same basic principles.

In practice, umbrella sampling simulations includes four steps. Step 1 is generating the sampling over the chosen reaction coordinate using restrained MD simulations. The reaction coordinate could be distance, angle, dihedral angle or other parameters which can be calculated and added harmonic restrain on in our simulations. The restrain force is added to generate sampling over the reaction coordinate. Step 2 is to generate the free energy profile based on the sampling over the reaction coordinate and eliminate the effect of the additional restrain force, usually using WHAM analysis. Then in step 3 we examine the output from the WHAM analysis. Step 4, we need repeating simulations to evaluate the precision.

1. **Do restrained MD simulation**
    a. decide on reaction coordinate range and window size
        i. this information is usually put in a DISANG file in sander, but might be in the mdin file if you are using RMSD as reaction coordinate

        ii. here is a sample DISANG file for dihedral center of mass restraints (note that this restraint uses a special version of sander). The r1, r2, r3 and r4 all need to be changed for each window. *r2 and r3 should always be the same value* (the target value for this window). rk2 and rk3 are the force constant and should always be the same value.

```
#  torsion restraint for phi of residue DG
 &rst  iat=-1,-1,-1,-1, r1=$low, r2=$i, r3=$i, r4=$high,
rk2 = 1000., rk3 = 1000.,

IGR1(1)=747,IGR1(2)=746,IGR1(3)=744,IGR1(4)=750,IGR1(5)
=753,IGR1(6)=754,

IGR1(7)=749,IGR1(8)=745,IGR1(9)=743,IGR1(10)=740,IGR1(1
1)=741,

IGR2(1)=735,IGR2(2)=755,IGR2(3)=757,IGR2(4)=738,IGR2(5)
=737,
```

```
IGR3(1)=768,IGR3(2)=788,IGR3(3)=790,IGR3(4)=771,IGR3(5)
=770,

IGR4(1)=773,IGR4(2)=774,IGR4(3)=776,IGR4(4)=787,IGR4(5)
=777,IGR4(6)=778,

IGR4(7)=786,IGR4(8)=779,IGR4(9)=780,IGR4(10)=782,IGR4(1
1)=783,IGR4(12)=276,

IGR4(13)=279,IGR4(14)=281,IGR4(15)=275,IGR4(16)=280,IGR
4(17)=273,IGR4(18)=271,
       IGR4(19)=270,
 &end
```

iii. here is a sample mdin file, it is suggested to use a script to create the files

```
md.in
 &cntrl
        imin = 0, ntx = 5, nstlim = 500000,
        ntc = 2, ntf = 1, tol=0.0000001, ntt = 1, dt =
0.002,
        ntb = 2, ntp = 1,
        ntwx = 500, ntwe = 0, ntwr = 500, ntpr = 500,
        scee = 1.2, cut = 8.0,
        ntr = 0, ibelly = 0, tempi = 300.0, temp0 =
330.0,
        nscm = 5000, iwrap = 1,
        nmropt = 1,
 /
 &wt type='DUMPFREQ', istep1=1  /
 &wt type='END'   /
DISANG=phi.RST.$ext
DUMPAVE=phi_vs_t.$ext
LISTIN=POUT
LISTOUT=POUT
```

b. decide on force constant (be careful of units for dihedral angles)
c. create initial coordinates for each window – two methods are possible
    i. run a single simulation with large force constant (probably larger than needed for the actual umbrella sampling runs) to move system through entire range, saving snapshots in traj file along the way, or saving into restart files (check sander manual). Take each restart file and run equilibration, then production dynamics (all with restraints at the target value of the window)
    ii. run a single simulation with restraints at the target value, and use the final coordinate from each window as the initial coordinate for the next window, changing the target value of the restraint
    iii. option (i) can be run efficiently in parallel, but (ii) may give the system more time to relax after the changes. Which you use will depend on how much and how quickly you expect the rest of the system to change in response to the changes induced by the

restraint.

d. Make sure that the initial coordinates for each window actually have reaction coordinate values that are in the range of the window. Increase the restraint force constant if needed and repeat the creation of initial coordinates.
e. Run the restrained MD simulations. Do a few of them first and then histogram the values from each window (from the DUMPAVE file).
    i. Make sure that the windows do not shift too far from the target values of the window (make sure the force constant is strong enough). You do not want there to be gaps in the profile because the restraints are too weak to keep the system at the high energy positions. If they move too much, increase the force constant. This is especially important where the energy is changing quickly (away from the minimum).
    ii. Make sure that the windows have enough overlap. This can be hardest to get when the system is near a minimum. If not, either decrease the force constant (may allow windows to shift away from desired value, see (i) above) or add extra windows (decrease the window spacing).

2. Calculate free energy profile using WHAM (check the manual and more info by Alan Grossfield at http://dasher.wustl.edu/alan/wham/)

a. make the metadata file, here is a sample (first line is a title, next lines have filename, then target value, then force constant (real force constant using V=1/2 * K * (r-r0)**2). *This is probably twice the value you put in the sander mdin since sander wants k/2, not k.*

```
#/path/to/timeseries/file    loc_win_min   spring
phi_vs_t/phi_vs_t.-090  -90     0.6
phi_vs_t/phi_vs_t.-085  -85     0.6
phi_vs_t/phi_vs_t.-080  -80     0.6
phi_vs_t/phi_vs_t.-075  -75     0.6
phi_vs_t/phi_vs_t.-070  -70     0.6
...
```

And so on until last window.

b. Run WHAM - here is a sample script for WHAM

```
#!/bin/csh -f

#hist_min and hist_max specify the boundaries of the histogram,
these need to be in the range in your metadata file
set hist_min = -90
set hist_max = 115

#number of bins in the histogram, and as a result the number
of points in the final PMF
set num_bins = 200
```

```
#convergence tolerance for the WHAM calculations
set tol = 0.001

#temperature in Kelvin at which the weighted histogram
calculation is performed
set temperature = 330

#number of "padding" values that should be printed for
periodic PMFs. This number should be set to 0 for aperiodic
reaction coordinates
set numpad = 1

#metadatafile specifies the name of the metadata file (has
file names for data from sander, target values and force
constants
set metadatafile = "PMF-metadata.in"

#freefile is the name used for the file containing the final
PMF and probability distribution
set freefile = "PMF.out"

#num_MC_trials and randSeed are both related to the
performance of Monte Carlo bootstrap error analysis - the
value you pick should be irrelevant, but I let the user set it
primarily for debugging purposes
set num_MC_trials = 10
set randSeed = 9999

/mnt/raidb/kensong/KIT/wham  $hist_min $hist_max $num_bins
$tol $temperature $numpad $metadatafile $freefile
$num_MC_trials $randSeed

# this is to shift the free energy profile range in this
dihedral example, probably should not be done for general use
awk '$1<=0{ print $1+360, $2, $3} $1>0{print $1, $2, $3}'
PMF.out > tmp
sort -g tmp > PMF.dat
```

3. **Look at your free energy profiles.** The first 2 columns are the data (reaction coordinate value and free energy), other columns give uncertainty info (check web site for more details).

4. **Determine the reliability of your data** by repeating with longer windows, or by creating new initial coordinates for each window, or by running more simulations to add to the data for each window (data from multiple simulations with the same target and force constant can be combined).

   Here is a sample complete script for generation of input files (dihedral angle center of mass restraints, run at teragrid):

```
------------------------------------------------------------
#!/bin/tcsh

set  i = -195
```

```
while ($i < 200)
#foreach i (-130 -110 -165 -175 115 140 180)

set curDir = $PWD

@ j = $i + 5

set ext = `printf "%3.3i" $i`

# Create one directory for each window
mkdir $curDir/WIND$ext
cd $curDir/WIND$ext

# Create the restraint input file
@ low = $i - 100
@ high = $i + 100

cat > phi.RST.$ext << EOF
#  torsion restraint for phi of residue DG
 &rst  iat=-1,-1,-1,-1, r1=$low, r2=$i, r3=$i, r4=$high,  rk2
= 1000., rk3 = 1000.,

IGR1(1)=747,IGR1(2)=746,IGR1(3)=744,IGR1(4)=750,IGR1(5)=753,IG
R1(6)=754,

IGR1(7)=749,IGR1(8)=745,IGR1(9)=743,IGR1(10)=740,IGR1(11)=741,

IGR2(1)=735,IGR2(2)=755,IGR2(3)=757,IGR2(4)=738,IGR2(5)=737,

IGR3(1)=768,IGR3(2)=788,IGR3(3)=790,IGR3(4)=771,IGR3(5)=770,

IGR4(1)=773,IGR4(2)=774,IGR4(3)=776,IGR4(4)=787,IGR4(5)=777,IG
R4(6)=778,
IGR4(7)=786,IGR4(8)=779,IGR4(9)=780,IGR4(10)=782,IGR4(11)=783,
IGR4(12)=276,IGR4(13)=279,IGR4(14)=281,IGR4(15)=275,IGR4(16)=2
80,IGR4(17)=273,IGR4(18)=271,IGR4(19)=270,

 &end

EOF

# Create the md input file
cat > md.in << EOF
md.in
 &cntrl
        imin = 0, ntx = 5, nstlim = 500000,
        ntc = 2, ntf = 1, tol=0.0000001, ntt = 1, dt = 0.002,
        ntb = 2, ntp = 1,
        ntwx = 500, ntwe = 0, ntwr = 500, ntpr = 500,
        scee = 1.2, cut = 8.0,
        ntr = 0, ibelly = 0, tempi = 300.0, temp0 = 330.0,
        nscm = 5000, iwrap = 1,
        nmropt = 1,
 /
 &wt type='DUMPFREQ', istep1=1  /
 &wt type='END'   /
```

```
DISANG=phi.RST.$ext
DUMPAVE=phi_vs_t.$ext
LISTIN=POUT
LISTOUT=POUT

END

EOF

# Create job file

set sander = "~kensong/amber9/src/sander.CMtorsion/sander.MPI"

cat > mdjob << EOF
#!/bin/csh


$sander \
-O \
-i md.in \
-o md.$ext.out \
-p ../../0GENE/sol.GC.parm7 \
-c ../../2INIT/WIND$ext/md.$ext.rst7 \
-x md.x  \
-e mden \
-v mdvel \
-inf mdinfo \
-r md.$ext.rst7


EOF

cat > runjobs << EOF
#!/bin/csh
#PBS -j oe
#PBS -l walltime=15:00:00
#PBS -l nodes=4:ppn=2
#PBS -r n
#PBS -M kensong
#PBS -N DW$ext
#PBS -V

set NP=\`wc -l \$PBS_NODEFILE | cut -d'/' -f1\`
cd \$PBS_O_WORKDIR

mpirun -machinefile \$PBS_NODEFILE -np \$NP mdjob
EOF

chmod u+x mdjob
chmod u+x runjobs

subJob:

qsub < runjobs

cd $curDir
```

```
echo $ext
@ i = $i + 5
end

-------------------------------------------------------------
```