Conformational Dynamics of the Bound and Unbound States of Human Alkyladenine Glycosylase

A Thesis submitted to the faculty of
San Francisco State University
In partial fulfillment of
the requirements for
the Degree

Master of Science

In

Chemistry: Biochemistry

by

Gabrielle Marie Garcia

San Francisco, California

January 2018

CERTIFICATION OF APPROVAL

I certify that I have read Conformational Dynamics of the Bound and Unbound States of Human Alkyladenine Glycosylase by Gabrielle Marie Garcia, and that in my opinion this work meets the criteria for approving a thesis submitted in partial fulfillment of the requirement for the degree Master of Science in Chemistry: Biochemistry at San Francisco State University.

Anton Guliaev, Ph.D.
Associate Professor

Misty Kuhn, Ph.D.
Assistant Professor

Andrew Ichimura, Ph.D.
Associate Professor

Conformational Dynamics of the Bound and Unbound States of Human Alkyladenine Glycosylase

Gabrielle Marie Garcia
San Francisco, California
2017

An intricate base excision repair (BER) mechanism maintains the integrity of human DNA against exposure to environmental pollutants and carcinogens, and natural error in DNA replication. Human alkyladenine glycosylase (hAAG) initiates BER for alkylated and deaminated adenine bases. The overall catalytic mechanism for hAAG is known, but the method of formation of the protein-DNA complex is not known. Two hypotheses for BER proteins compare a sliding vs. clamping mechanism to search for lesions. In this work, conformational differences between bound and unbound enzyme were studied via molecular dynamics (MD) simulations. Analysis was based on MD trajectory data and included calculations of hydrogen bond interactions, β-factor, and RMS fluctuations. We propose hAAG searches the DNA duplex via sliding mechanism. The insights obtained in this work point to the development of future cancer therapeutics.

I certify that the Abstract is a correct representation of the content of this thesis.

Chair, Thesis Committee

12/18/2017
Date

PREFACE AND/OR ACKNOWLEDGEMENTS

This has been an accumulation of many years' work and would not have been possible without the support of certain academic peers, supervisors, family members and friends. First and foremost I would like to thank Dr. Anton Guliaev. Thank you for your vast amount of patience as I learned the ropes of the lab. I appreciate your willingness to be a mentor to me and that you were always flexible with my demanding schedule. I cannot forget to thank Dr. Baird who may not realize the impact he had on my academic career. Thank you, Dr. Baird, for the countless times I came to you for advising. You pointed me in the direction of joining a lab as an undergraduate and encouraged me when I doubted my potential. A big thank you to my family and friends who have never stopped believing in me and supporting me through the years.

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figures Page

LIST OF APPENDICES

LIST OF EQUATIONS

Equations Page

Cancer is pervasive and is a disease without boundaries—It affects young and old, and people of every nationality. Of the many types of cancer, colorectal and liver cancers are among the leading causes of death in the world. According to the 2015 World Health Report, there were over 8.8 million deaths caused by cancer—788,000 and 774,000 of which were caused by liver and colorectal cancer, respectively. This is a 1% increase from the 2014 report.[1] There are a variety of reasons why people develop cancer, some known and some unknown. The rates of liver and colorectal cancers appear only to be increasing as alkylating agents, such as nitrosamine, vinyl chloride[2] and polyurethane, are encountered within the environment at elevating amounts.[3] Exposure to tobacco products and environmental pollutants has long been known to cause a variety of health issues; however, these specific pollutants are detrimental to DNA integrity as they are IARC-classified carcinogens which are linked to liver and colorectal cancers.[4,5] Exposure to environmental pollutants occurs primarily in and around factories, but also through personal use of tobacco products, such as smoking or "chewing". Nitrosamines in tobacco products damage DNA via alkylation.[6,7] Although there are regulations within the US that limit how much urethane or vinyl chloride is released from factories, there aren't such regulations everywhere these environmental pollutants are found around the world. Not all countries have strict regulations that protect public health, such as where it is acceptable to smoke tobacco products, or put limitations on the amount of vinyl chloride which is released into the air from factories producing and handling plastics.[8] Nitrosamines and urethanes have been labeled as contributors to the rising liver and

colorectal cancer rates through DNA alkylation as well.[9] These alkylating agents are known to mutate DNA bases, producing hypoxanthine, 1,N^6-ethenoadenine, and 3-

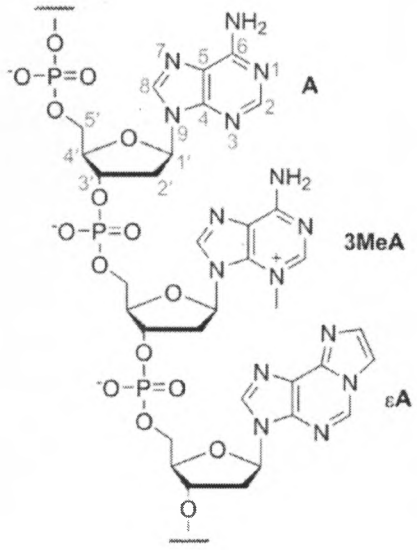


Figure 1. DNA backbone with the following three bases: adenine (A), 2-methyladenine (3MeA), and 1,N^6-ethenoadenine (εA).[3]

methyladenine (Figure 1)[3].[10] Moreover, in some cases patients are intentionally exposed to extremely high concentrations of alkylating agents during chemotherapy for various cancers, including glioblastoma, lymphomas, melanoma and leukemias. The good news is the body has its own defense mechanism to repair lesions which result from both carcinogens or mistakes in DNA replication.[11] This mechanism is known as base excision repair (BER).[12] DNA glycosylases initiate the process of BER with the removal of damaged DNA bases by glycosidic bond cleavage.[13] Once a damaged base is removed and an AP-site is exposed, endonuclease and ligase insert the correct base in place and

seal the DNA backbone.[14] The glycosylase employed in BER is dependent upon the type of lesion involved. One specific glycosylase, which is the focus of this thesis, is human alkyladenine glycosylase (hAAG). It is the only DNA glycosylase in human cells to

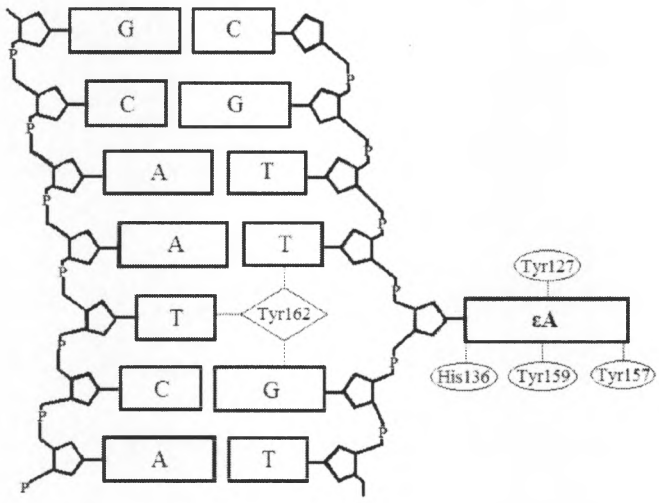


Figure 2. The hAAG probe, Tyr162 stabilizes the DNA structure while an ethenoadenine (εA) lesion is held in the active site pocket by pi-to-pi stacking with Tyr127, His136, Tyr159 and Tyr157.[18]

remove highly mutagenic alkylated purine bases, such as 1,N6-ethenoadenine (εA), 3-methyladenine, and 7-methylguanine. hAAG specifically targets alkylated and deaminated purines, including the modified adenine bases, such as the three mentioned previously. For mutations involving adenine bases, the active site within hAAG executes glycosidic bond cleavage through the following steps: Residue Y162 acts as a probe that inserts into the DNA, flipping the DNA lesion out and into hAAG's active site pocket.[15] The lesion is then held in place by residues H136, Y159, Y157, and Y127 through π-π stacking.[16,17] This allows successful nucleophilic attack of a deprotonated water on the C1' side of the substrate ribose (Figure 2) [18]. Once the lesion is cleaved, the base is

removed and polymerase moves in with the correct base. Ligase then seals the DNA backbone ending the process of BER.[19]

Since glycosylases play a significant role in initiating BER, it is vital to this area of cancer research to understand mechanisms of lesion recognition by hAAG and if there are any ways by which the efficiency of such mechanism can be improved. The catalytic mechanism for hAAG has been stablished based on the crystallographic data for the hAAG-DNA complex containing DNA adduct. However, there is no crystal structure of the wild-type hAAG without bound DNA (unbound hAAG or "apo form").[18] The lack of the structural information for the unbound form of hAAG prevents from obtaining important insights into formation of protein-DNA complex and initial lesion recognition by this repair enzyme. The conformational data of the unbound form, together with current data on the hAAG-DNA complex will be essential for the understanding how this enzyme search for the sites of DNA damage. Improvement of hAAG's efficiency in DNA repair can be important for reducing liver and colorectal rates than regulation of environmental pollutants alone. Since these cancers are difficult to treat and tumor removal may cause physical and emotional trauma, it is best to determine preventative methods. Regardless of governmental regulations or the human body's ability to excise these carcinogenic DNA lesions, liver and colorectal cancer rates continue to grow throughout the population. Structural details on the unbound hAAG would provide better understanding of the mechanisms of lesion removal. Such research could aid in improving hAAG's base excision repair efficiency.

The overall goal of this research project is to use molecular dynamics simulations to elucidate the structure and dynamics of hAAG without the DNA present. The conformational dynamics of hAAG will provide valuable insights on the protein behavior prior to the formation of protein-DNA complex. The obtained data will be compared with structural data previously obtained by x-ray crystallography on hAAG-DNA complex. The structural differences between unbound hAAG and hAAG complexed with DNA should point to essential features leading to formation of the complex and recognition of the damaged sites. This project will aid in future development of cancer therapeutics and prevention.

Theory of Molecular Dynamics

1. Introduction to Molecular Dynamics

As technology continues to advance the methods used in today's typical non-computational research lab, many scientists are turning to computational chemistry to aid in describing or studying a system, whether at a micro- or macroscale. The development of molecular dynamics (MD) first began in the early 1950s by physics researchers, Metropolis et. al. They published "Equation of State Calculations by Fast Computing Machines", in which they discuss the foundational physics of today's MD programs.[20] The first article published on protein folding was in 1975 by Levitt and Warshel.[21] In 2013 they went on to win the Nobel prize along with Karplus "for the development of multiscale models for complex chemical systems."[22] They showed that MD simulations can be used to explain chemical phenomena at an atomic level. Computational chemistry also makes it possible for scientists to study the intricate details of a protein at atomic and molecular levels without physically entering a 'wet lab' or when experimental data cannot be obtained.

Current applications of MD include prediction of protein folding, prediction of enzyme mechanism/function, calculations of binding energies, calculation of potentials for interfaces within a polymer, etc.[23]

Before taking a look at the physics behind the calculations used within a MD simulation, it is important to have a well-defined starting structure. Most MD capable programs, such as AMBER, allow users to build a structure by inputting a specific

sequence—or, by submitting coordinates of a known structure determined by x-ray crystallography.[24] The x-ray data is stored in the form of a PDB file. PDB files are commonly referenced in research articles and are easily obtained on RCSB.org (Research Collaboratory for Structural Bioinformatics). They are essentially maps for the protein/molecular structure one may want to simulate. Once a structure is obtained there are preliminary steps to execute before data collection begins. Since proteins are not static assemblages, they must first be put back into a phase which simulates the natural, biochemical environment for the protein. This is done through neutralization, hydration, energy minimization and equilibration. Neutralization allows the protein to be studied without artificial charges interfering with MD calculations. The addition of counter ions neutralizes the system to prevent unwanted electrostatic interactions. Hydration is typically achieved by the addition of box with explicit water molecules ("water box"). The size of the box is larger than the size of the protein being studied. It is a good practice to have at least 8Å of solvent with counter ions around the protein .The initial energy minimization reduces the energetic tensions imposed on the structure during crystallization. Equilibration allows the system to slowly come to room temperature while the water box achieves standard density of 1.01g/cm3. These preliminary processes are all done using AMBER Tool, Leap. Leap is used specifically for preparing input files for simulation. The series of short minimization and MD runs are used to equilibrate the explicit solvent together with the protein. Now that the preparation has been reviewed, the theory which supports data collection in MD simulations will be discussed.

Most of the theory of MD simulations actually comes from physics. Most specifically, Newton's second law is a fundamental theory of how MD programs have been built. MD simulations track the position, acceleration, energy with respect to change in time and under specific conditions. For example, a condition of the simulation may be the solution present or which force field is used to model simulation MD simulations, depending on how complex the molecule or protein and its environment, can take up a lot of computational space and time. Thankfully, Newton's second law is one easy foundational equation which helps explain how MD simulations work.

Figure 3. Spring model for two atoms, A and B.
k_s is the spring constant from Hooke's law.

$$\frac{d^2x_i}{dt^2} = \frac{F_{x_i}}{m_i} \qquad (1)$$

Equation 1 describes how changes in position and velocity can be determined for an atom with a specific mass under a given force. In simplest terms, it is written as F = ma. One way this can be modeled is with displacement for the diatomic system. The relationship between potential energy and force is seen in the diatomic spring model.

The displacement for any given diatomic system can be described given the relationship described in Hooke's law. Equations 2a-e describe the process of deriving the force as it relates to displacement between two atoms A and B (figure 3) as a function of the force found between the two atoms. This force is equal to the potential energy gradient.

$$F_S = -k_S(R - R_e) \tag{2a}$$

$$\frac{dU(x)}{dx} = -F \tag{2b}$$

$$U(x) = -\int F(x)dx \tag{2c}$$

$$U(R) = U(R_e) + \frac{1}{2}k_S(R - R_e)^2 \tag{2d}$$

$$U(R) = \frac{1}{2}k_S(R - R_e)^2 \tag{2e}$$

Equations 3a-d describe the relationship between the force and position of atoms in a diatomic system by relating the force found in equations 2a-e to Newton's second law.

For atom A:

$$Fa = m_1\frac{d^2x_1}{dt^2} = k_S(R - R_e) = k_S(x_2 - x_1 - R_e) \tag{3a}$$

For atom B:

$$m_2\frac{d^2x_2}{dt^2} = -k_S(R - R_e) = -k_S(x_2 - x_1 - R_e) \tag{3b}$$

$$(m_1 + m_2)\frac{d^2R}{dt^2} = -k_S(R - R_e) \tag{3c}$$

$$\frac{d^2R}{dt^2} = -k_S(\frac{1}{m_1} + \frac{1}{m_2})(R - R_e) \tag{3d}$$

When the computations are being made for individual atoms in MD, each position and change in velocity is stored in a file called a trajectory file. These trajectory files are saved with respect to a certain time-step. In order to get the bigger picture of how these consecutive changes work as a whole, it is important to look at the trajectory. Before these calculations can be done, it is important to establish which force is to be applied.

2. Force Models and Force Fields

In 1957 Adler and Wainwright investigated the first type of MD force model to be used. They followed Newton's first law which states that a particle has a constant velocity and that the change of position between two collisions is defined as $v_i \partial t$. Another model shows force as constant between two collisions. This means that the direction and velocity for a particle would remain constant until it collides with another particle. Motion cannot be most accurately described using just these two models, and so Rahman used a model in 1964 to simulate Argon, which relied on the idea that the force on one particle is dependent upon the position and force on another particle. This can be thought of as a domino effect and points to the idea that molecular motions are fluid and dynamic. With the intertwining of many movements in a system, individual motions cannot be solved analytically without using integration of equations for the finite difference method.

In order to describe the force on each atom, the differential equation for potential energy is given by:

$$\vec{F}_j = -\vec{\nabla}_j U(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_n) \qquad (4)$$

This characterization of the positions and velocities by molecular mechanics does not take into account the electronic motions the way that quantum mechanics does. There are additional qualities which need to be acknowledged because they contribute to the potential energy found in the system. Bonded and non-bonded interactions between atoms can be represented arbitrarily as contributing to the total energy of a system.

$$E_{total} = E_{bonded} + E_{nonbonded} \qquad (5)$$

Motions attributed to bonded atoms include stretching, torsion (bond rotation), and angular bonding. As mentioned previously, non-bonded interactions are significant because they contribute to the motions and positions of the surrounding atoms. Motions attributed to non-bonded atoms are a result of electrostatic affects and van der Waals interactions (figure 4).[25]

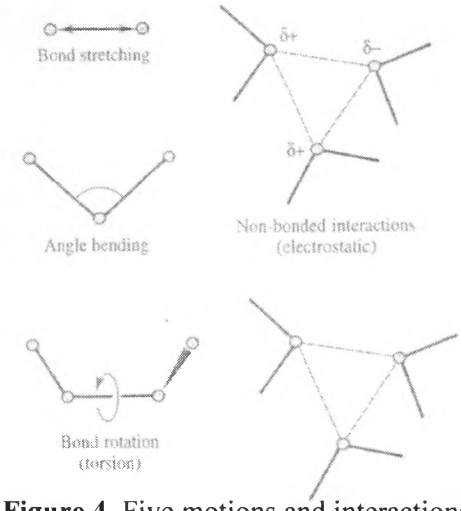


Figure 4. Five motions and interactions which contribute to total potential energy.[25]

As one can see, there is much more to the total energy than just what's described in equation 5. The detailed expansion of this total energy equation is written as a function of positions (r) for N particles.

$$E(\vec{r}^N) = \sum_{bonds} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{torsions} \frac{V_n}{2}(1 + \cos(n\omega - \gamma))$$

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

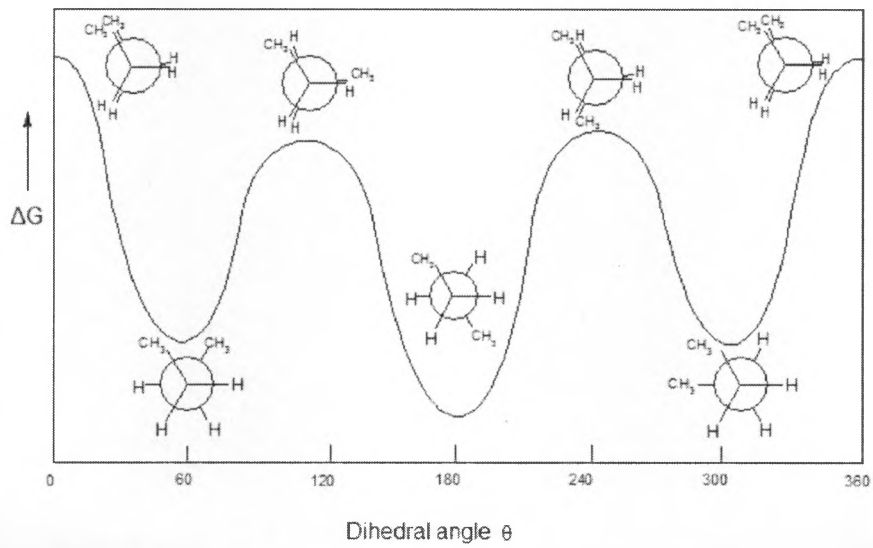


Figure 5. Butane is provided as an example of how the potential energy changes as the angle changes between the end methyl groups.[26]

Hooke's Law describes the summation of the stretching and bond angle and how potential energy changes with respect to bond length (l) and bond angle (theta) as they deviate from the reference value, l0. Bond rotations, defined as torsion, are also included because of the implicit effects of these bonded interactions on energy. Figure 5[26] below

shows the significant impact that torsion angle has across various energy conformations.[27]

3. Using Finite Difference to Predict Change in Position

With the calculated potential energy gradient, a force and acceleration is calculated for the simulation. Integration of the acceleration, with respect to time, results in a new velocity. Further integration gives the change in position (distance). Once the new positions are determined, the potential energy is calculated again to provide the new force and acceleration given for the current positions.

From $F_i = m_i a_i$,

$$m_i\left(\frac{d^2 r_i}{dt^2}\right) = -\frac{dU}{dr_i} \tag{7}$$

Since acceleration is defined as the change in velocity and velocity is the change in position, both with respect to time, final changes in position can be calculated via integration as follows:

$$\int a dt = \int_{v_0}^{v_1} dv \tag{8a}$$

$$a(t_1 - t_0) = v_1 - v_0 \tag{8b}$$

$$v_1 = v_0 + a(t_1 - t_0) = v_0 - a\Delta t \tag{8c}$$

$$v = \frac{dr_i}{dt} \tag{8d}$$

$$\int v dt = \int_{r_0}^{r_1} dr \tag{8e}$$

$$v_1(t_1 - t_0) = r_1 - r_0 \tag{8f}$$

$$r_1 = v_1(t_1 - t_0) + r_0 = v_1\Delta t + r_0 \ldots \tag{8g}$$

As mentioned before, Newton's second law is the foundation for a multitude of calculations used in molecular dynamics simulations. The overall method is known as the finite difference method and is found in a variety of algorithms. The idea is that with a constant total force acting on all atoms in a system over a given period of time, the changes in acceleration can be calculated. This allows for the prediction of new positions for each atom in the system with respect to change in time. The finite difference is used across many different algorithms. A common algorithm used is the Verlet algorithm.[28] The Verlet algorithm (equation 9a) does have its disadvantages though because it doesn't calculate velocities and may have a smaller precision than other algorithms. The lack in precision may be the result of adding the smaller acceleration term to two larger terms. Velocities can be found using another algorithm (equation 9b).

$$\vec{r}(t+\delta t) = 2\vec{r}(t) - \vec{r}(t-\delta t) + \frac{1}{2}\delta t^2 a(t) + \cdots \qquad (9a)$$

$$\vec{v}(t) = [\vec{r}(t+\delta t) - \vec{r}(t-\delta t)]/2\delta t \qquad (9b)$$

The Verlet algorithm has been developed into other algorithms which improves the precision and includes calculating the velocities. One such algorithm is the leap-frog algorithm. The name comes from the fact that the positions and velocities are not calculated simultaneously—one is calculated to determine the other and so forth. The leap-frog algorithm was developed and used with AMBER programs Sander and PMEMD.[29] The velocities are first calculated in equation 10a. The velocities with respect to time t, and positions are then calculated in equation 10b and equation 10c.

$$\vec{v}(t+\tfrac{1}{2}\delta t) = \vec{v}(t-\tfrac{1}{2}\delta t) + \delta t\vec{a}(t) \quad (10a)$$

$$\vec{v}(t) = \tfrac{1}{2}[\vec{v}\left(t-\tfrac{1}{2}\delta t\right) + \vec{v}(t-\tfrac{1}{2}\delta t) \quad (10b)$$

$$\vec{r}(t+\delta t) = \vec{r}(t) - \delta t\vec{v}(t+\tfrac{1}{2}\delta t) \quad (10c)$$

Additional algorithms which are commonly used are the Beeman's algorithm[30] and the Taylor series expansion, which are described in equations 11a-c.

$$\vec{r}(t+\delta t) = \vec{r}(t) + \delta t\vec{v}(t) + \tfrac{1}{2}\delta t^2\vec{a}(t) + \tfrac{1}{6}\delta t^3\vec{b}(t) + \cdots \quad (11a)$$

$$\vec{v}(t+\delta t) = \vec{v}(t) + \delta t\vec{a}(t) + \tfrac{1}{2}\delta t^2\vec{b}(t) + \tfrac{1}{6}\delta t^3\vec{c}(t) + \cdots \quad (11b)$$

$$\vec{a}(t+\delta t) = \vec{a}(t) + \delta t\vec{b}(t) + \tfrac{1}{2}\delta t^2\vec{c}(t) + \cdots \quad (11c)$$

In addition to describing the motions and positions of each individual atom in simulation, it is vital that the program being used to calculate these algorithms has an accurate description of the character or type of each atom, bond and angle involved. This is where the importance of an accurate source comes into play when initially setting up a model for molecular dynamics. The program used for this thesis is known as AMBER, which stands for Assisted Model Building with Energy Refinement. AMBER has multiple packages within its program which host an array of libraries. These libraries contain information about the force field that is to be used according to which atom types are included. Even so, the atom types are assigned by the libraries which include information

about how certain atoms are connected and how they interact with other molecules, atoms, forces, charges, etc. The force field used will be further described in the methods section of this composition.

4. Introducing Solvation Methods of MD

There are times when studying a molecule or system in a vacuum is important. For the sake of this project it is imperative that the solutions dynamics and interactions be understood. This project works with an enzyme, and as with any biomolecular study, there will be fluids involved. Intermolecular interactions play a significant role in how certain biomolecules such as enzymes function. Human alkyladenine glycosylase, the main focus of this project, even employs the use of a water molecule to help execute glycosidic bond cleavage of a mutated adenine base from DNA. There are two ways which will be used to describe the solvent in an MD simulation: explicit and implicit solvation. Both solvent methods have their advantages and disadvantages and are used depending upon the information that is needed about the system being simulated. With larger molecular systems, using explicit solvation requires more time for calculations because of the vast quantity of water atoms which will be included in the velocity, position and acceleration output. In some cases it is too expensive computationally to run a simulation with explicit solvation—implicit solvation may then be the better choice.[31,32]

As shown in figure 6 below, implicit solvation is named appropriately, as the presence of water molecules and their properties are described as average values (rather than individually).[33] This implication of an average solution comes with its cost as well. Using implicit solvation brings to question whether an average force field can be expressed without consideration of the solvent degrees of freedom as done in the explicit solvent model. In order to check the consistency in expression between the two models, the potential mean force is used to configure the probability distribution function, which is based on a system in equilibrium.[34]

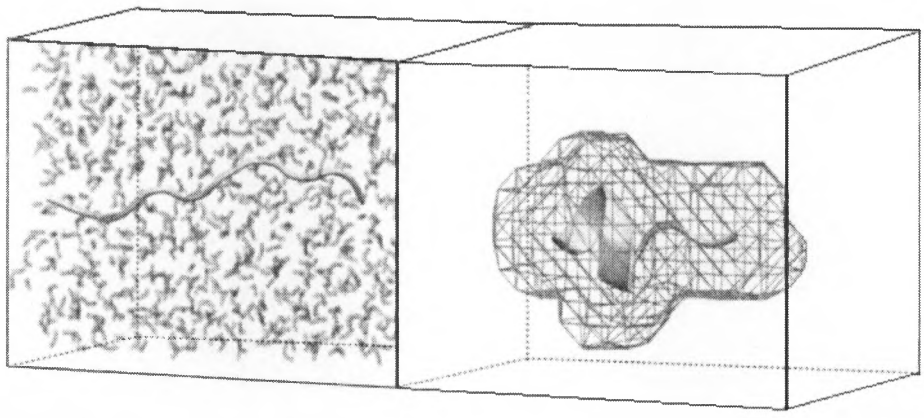


Figure 6. Individual water molecules in explicit solvation (left) vs distribution of potential energy distributed around solute in implicit solvation (right).

Reducing the probability function down to this expression allows for separation of solvent/solute representation—the solvent can be represented implicitly while the solute is studied explicitly. This is clearly an advantage in that the description of the solvent is simplified, allowing for more data analysis and computational space regarding the target

molecule, the solute. With the focus on the solute and off the details of solute-solvent interactions, this may be a disadvantage. Only having non-specific interactions between solvent and solute may mean that ionization effects have been ignored and electrostatic energy may be misinterpreted. Solvation of the solute is done before minimization and equilibration.

5. AMBER and Implicit Solvation

AMBER, Assisted model building with energy refinement, uses implicit solvation methods known as the general Born (GB) solvation model. Although implicit solvation was not used in this project, it is a computational program used for data acquisition in other types of simulation projects. The GB model supposes that the total solvation free energy of a molecule can be broken down into two parts, electrostatic and non-electrostatic. These are represented in the expression $\Delta G_{solv} = \Delta G_{el} + \Delta G_{nonel}$. Electrostatic interactions represent the free energy from first removing all charges in the vacuum and adding them back in the presence of a continuum environment.[35,36] The non-electrostatic interactions come from favorable van der Waals interactions between solute and solvent molecules. AMBER's methods of calculating ΔG_{el} have been considered efficient when compared with other models such as the Poisson-Boltzmann model. AMBER approximates each individual atom as a sphere with a radius R and charge q. A dielectric constant of 1 is assigned and is based on the assumption that the interior of each atom is uniformly filled. For water at 300K, molecules are surrounded by a solvent with a

high dielectric of 80. This allows the approximation of ΔG_{el} as described in equation 12, where r_{ij} is the distance between atoms *i* and *j*, R_i and R_j are the effective Born radii, f_{GB} is a smooth function of the contained arguments. The Debye-Huckel screening parameter, κ, provides a means for incorporating electrostatic effects of salt.

$$\Delta G_{el} = -\frac{1}{2}\sum_{ij} \frac{q_i q_j}{f_{GB}(r_{ij},R_i,R_j)}\left(1 - \frac{\exp[-\kappa f_{GB}]}{\varepsilon}\right) \qquad (12a)$$

The parameter f_{GB} is typically implemented as $f_{GB} = r_{ij}^2 + R_i R_j \, exp\left(\frac{-r_{ij}^2}{4R_i R_j}\right)^{1/2}$. For isolated ions, the Born radius can be approximated as equal to its van der Waals radius, ρ_i. If the assumption is made that $\kappa = 0$ in pure water, then ΔG_{el} can be simplified further to give equation 12b.

$$\Delta G_{el} = -\frac{q_i^2}{2\rho_i}\left(1 - \frac{1}{\varepsilon}\right) \qquad (12b)$$

6. AMBER and explicit solvation

For the course of this project, explicit solvation has been used to describe water molecules in MD. Explicit solvation involves describing the position and effects of each individual solvent molecule. Although it is more computationally expensive to describe the solvent explicitly, it is far more accurate. This is due to the fact that instead of estimating the positions and effects of water molecules as an average through the space of the simulation, each individual water molecule is tracked and accounted for. Of course, in order to accurately describe any solute-solvent dynamics or how the solvent behaves near the edge of the vacuum's boundary, periodic boundary conditions are set. Periodic

boundary conditions put the solute in a box that is replicated in three dimensions for the course of the simulation. To cut down on calculations, only one solvent molecule is specifically described, but the effects and behavior of such a solvent molecule is reproduced over the rest of the simulation in other three-dimensional boxes. This approach is only beneficial if there is a non-bonded cutoff distance put in place. Without a non-bonded cutoff distance, the calculations become computationally expensive. This is because the computer will attempt to calculate any residual effects of molecules that are not even capable of interacting because of the distance between them.

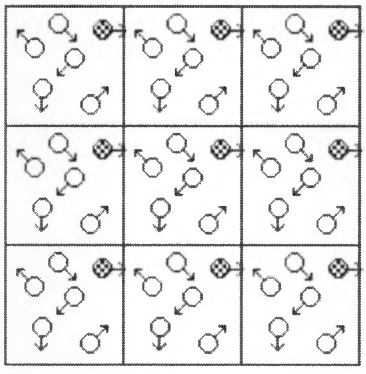


Figure 7. Periodic boundary conditions in two dimensions.

The cutoff distance in AMBER cannot be longer than half the length of the solvent box dimension. There are different methods of modeling water molecules within a simulation using explicit solvent model. Atomic charges, Lennard-Jones parameters and geometry all play significant roles in how a water model is defined.

7. Minimization and Equilibration for the Initial Structure

When simulating a molecule, the potential energy gradient and starting force are calculated given the atoms and bonds described in the Protein Database (PDB) file. The PDB acts as a map of the molecule and is typically the crystal structure. Since biological processes are not executed in a crystal state, it is important for the integrity of the study that the structure for simulation is prepared in a way that mirrors the actual biological condition it may be in naturally. What this entails is processing the PDB file containing the crystal structure through multiple simulations. The first is energy minimization of the structure. Energy minimization adjusts the coordinates with a modified conjugate gradient minimizer until the energy function is at a relative minimum. This can be a problem due to long-range interactions—however, there is a cutoff range that is typically set to exclude interactions past the range of 8-10Å.[37] When the structural strains from conflicting charges and non-ideal bond angles are minimized, the structure goes from a rigid state to a relaxed state. In some cases, counter-ions may be added to the system so that the simulation is run in a neutral state. If counter-ions are added, it would be done so before minimization is run.

Equilibration is somewhat like minimization in that it relaxes tensions or strains within the structure—however, equilibration is needed to further 'melt' the solid lattice structure that was provided in the PDB, allowing the even more 'relaxed' state to come forward. In this process, parameters are set for a specific pressure or temperature and the system being simulated may have the pressure held constant while the temperature is

slowly brought up to a set degree Kelvin. For most biological studies, the temperature set for equilibration is 310K which is approximate to body temperature. For simulations involving solvation with a water box, it is also important to set a parameter so that the water density matches the average value of 1.01 g/cm^3. The specifics of the minimization and equilibration parameters used for this thesis will be further discussed in the methods section.

8. Running an MD Production

As defined by Merriam-Webster, a trajectory is the path followed by a projectile flying or an object moving under the action of given forces. In the realm of computational chemistry and for the scope of this project, a trajectory is the accumulation of conformational snapshots occurring in consecutive order for the myriad of atoms provided in the system. When given specific time intervals, the equations calculate acceleration, velocity and finally a new position for every atom described in the simulation. Each calculation is based on a starting state resulting in a final state. The final state is then used as the starting state in the new calculation. Every time a final state is calculated it can be thought of as a snapshot of data. These snapshots make up what can be compared to frames in a movie and the consequential accumulation of these frames is known as a trajectory. This means for each calculation explained earlier in this paper, one position or velocity is recorded and is saved as a trajectory file. In order to get the overall scope of how the system changes with time, hundreds of thousands of snapshots

(trajectories) of the atomic positions need to be accumulated. The trajectory files can then be processed to extract different types of data, including the RMSd of the positions for each residue, the b-factor, the potential energy changes, etc. From this perspective, running a MD production to gather trajectory files sounds simple—however, there are even finer details to setting up a trajectory which are necessary for the sake of accuracy. Parameters that may be set include, but are not limited to, the temperature, pressure, length (how many calculations) and step size (how often). Temperature, pressure, and even length make sense as these are parameters that are often recorded in any given lab experiment. In computational chemistry or with molecular dynamics simulations, step-size is a very important factor to consider when setting up any production run.

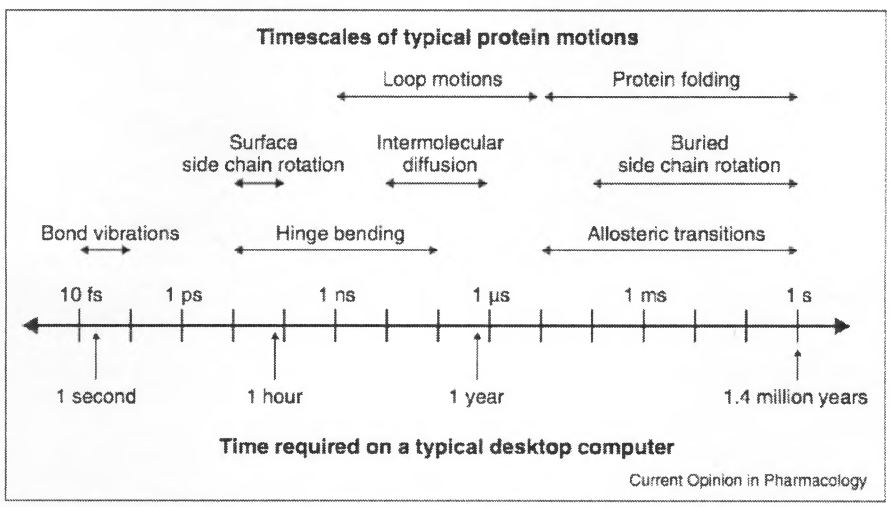


Figure 8. Timescales of typical protein motions from the atomic level to macromolecular reorganization.[38]

The step-size will allow for the user to understand processes or movements in a system at varying degrees of observation. As shown in figure 8, loop motions occur in the range of

1 ns to 1 μs, while bond vibrations are seen at femtoseconds.[38] When studying a protein and its conformational dynamics, it is appropriate to set a time-step for each trajectory calculation at about 2 femtoseconds. If a larger time-step were set then certain motions might be missed and the character of the enzyme could be misinterpreted. This is within range of what Jo and Kim found to be accurate for atomic motions. They determined a range of 1.8 to 3.8 femtoseconds to be accurate for time-steps involving atomic motions.[39] In addition to choosing a time-step, it is vital to the accuracy of the analysis to choose a simulation time long enough. For example, loop motions may be observed between the nanosecond and microsecond timescale. For the sake of accuracy, it is best to gather data for a production run on the microsecond timescale.

In addition to considering the motions which are being focused on, it is also important to be sure that the simulation time is long enough to take the system through all of its possible ensembles.[40] Ensembles refer to an average taken over a large number of replicas of the system considered simultaneously. The equation for ensemble average is seen in equation 13a,

$$\langle A \rangle_{ensemble} = \iint dp^N dr^N \, A(p^N, r^N)\rho(p^N, r^N) \qquad (13a)$$

where $A(p^N, r^N)$ is the observable of interest and is expressed as a function of the momenta (p) and the positions (r) of the system. The integration is taken over all possible variables of r and p. The probability density, $\rho(p^N, r^N)$, is given by equation 13b, where H is the Hamiltonian, T is temperature and k_B is Boltzmann's constant, and Q is the partition function.

$$\rho(p^N, r^N) = \frac{1}{Q} exp[-H(p^N, r^N)/k_B T] \qquad (13b)$$

These calculations are extremely involved and must be calculated for all possible states within the system. In MD, the Ergodic hypothesis is followed and states the ensemble average is equal to the time average. What the Ergodic hypothesis means is that if the system is allowed to evolve in time indefinitely, the system will eventually pass through all possible states. This is why in MD it is important to generate enough representative conformations such that this equality is satisfied. The time average is provided in equation 14, where M is the number of time steps in the simulation and $A(p^N, r^N)$ is the instantaneous value of A.

$$\langle A \rangle_{time} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A(p^N(t), r^N(t)) dt \approx \frac{1}{M} \sum_{t=1}^{M} A(p^N, r^N) \qquad (14)$$

This small time-step of 2 femtoseconds accumulated over microseconds of data means there are millions of calculations being made for each individual atom in simulation (500,000,000 steps). In the project of this thesis, macromolecular shifts are being questioned and observed, so the time-step and timescale chosen was also 2 femtoseconds and microseconds, respectively. This is a vast amount of data to sort through and often requires the use of a visual software to help the user understand what is happening with the system throughout data accumulation.

In this thesis, the trajectory data is visually analyzed with a program known as Visual Molecular Dynamics (VMD). VMD takes the atoms, positions and velocities calculated and visually presents them in an almost tangible way, allowing users to click to identify

an atom, rotate the structure and zoom in on certain parts of the protein. VMD's capabilities are vast and include calculating root-mean-square-deviations of changes in distance from the starting structure and throughout the trajectory.[41] It can be used to align similar structures and in this study, to characterize the conformational dynamics of hAAG. Visual analysis may be enough to make predictions about a protein's dynamic or give hint to its mechanism—however, visual analysis is not typically enough as a supporting argument in a new finding. Therefore, MD programs are even more useful.

Programs like AMBER come with a set of tools which are designed to computationally analyze the data output from a production run. Some methods of analysis include, changes in H-bonds found within the trajectory, flexibility or mobility of macromolecular structures and individual residues, charge distribution, etc. The main AMBERTools used in this project are tLEaP and CPPTRAJ. LEaP is used for preliminary data preparation and CPPTRAJ processes trajectories collected at the end of an MD simulation. The individual uses for each program will be elaborated in the experimental methods section.

Experimental Methods

Human alkyladenine glycosylase is an enzyme which has been well-studied, yet the wild type crystal structure has yet to be determined. The active-site mechanism has been outlined and confirmed with various studies, but only while hAAG is complexed with lesion-containing DNA. MD simulations prove beneficial in this situation because the DNA can be manually removed from the known crystal structure. This allows researchers to get a more complete picture of hAAG's conformational changes in its wild type form. With this knowledge, the overall mechanism and behavior of hAAG will be further understood and research may be done to improve the efficiency of BER of adenine lesions by hAAG. For this project, Assisted Model Building with Energy Refinement (AMBER) is the program that was used to evaluate conformations of unbound hAAG. For this thesis, there are two simulations for which the experimental methods will be described. The first one is a simulation of hAAG in its wild type form. The second includes hAAG complexed to 12-mer DNA containing εA lesion.

1. Choosing and Preparing the Structure

The initial structure used for this thesis comes from the RCSB database PDB 1F4R. PDB 1F4R is a crystal structure of hAAG-DNA complex with a single εA lesion incorporated across from a thymine base within the DNA. [42] The final structure was determined by Lau et. al. using Crystallography and NMR System (CNS).[43]

For the first simulation (simulation I), the 1F4R PDB was modified to remove data pertaining to the DNA molecule. This left only the atomic descriptions of hAAG in its wild type form. The second simulation (simulation II) didn't require any modifications to the crystal structure because it contained hAAG-DNA complex with an εA lesion present.

For AMBER to be able to calculate production run data for each simulation, there are a few additional files needed other than the structure itself. The PDB coordinates are processed to generate parameter and coordinate files in unique AMBER format . These are often referred to as *parm7* and *crd* files. AMBER needs more than just the atoms coordinates. This is where the *parm7* file is vital. The *parm7* is sometimes referred to as the topology of the simulation, and tells AMBER how the atoms in the *crd* file are connected or bonded. It also contains important information regarding force-field parameters, such as atom charges, atom types, bond and angle force constants, equilibrium values and etc. These two files don't come automatically prepared with the

PDB from the RCSB webpage. They are generated by AMBER's tool called LEaP (Link, Edit, and Parm). LEaP contains a library with information about standard nucleic acids and various atom types. The information contained within the PDB file is used in conjunction with its LEaP's library to generate the *parm7* and *crd* files.

The simulation for the bound conformation of hAAG may not have required any modifications to the PDB structure—however, there is more extensive work to be done in order to generate parameter and coordinate files due to the εA lesion present. As mentioned previously, LEaP contains libraries of information for standard residues and atom types. Since εA is not a standard residue, a separate library and parameter file must be generated for εA using another AMBER tool known as ANTECHAMBER. What ANTECHAMBER does is take the given structure for a known molecule such as εA and generate a parameter file for that individual molecule. This allows the user to then take the parameters of εA and input them to LEaP so that when *parm7* and *crd* files are being generated, it knows how to read or interpret the structure for εA.

For both simulations, the remainder of the preparation and processing is the same. The next step in preparing for the first set of simulations is addition of counterions. When hAAG, or any biological matter, is crystallized, it is being dehydrated and placed under specific conditions needed for successful crystallization. Therefore, it is important to represent physiological setting to avoid errors in the simulation. One way this is done is through the addition of counterions. There are multiple studies which suggest a negative impact of high-salt concentration. Studies have indicated that protein and protein-DNA

complexes benefit and are further stabilized when counterions are included.[44] When counterions are added to the simulation, the result is a neutral molecule or complex. The overall charge on hAAG or hAAG complexed with DNA is negative and therefore required the addition of sodium ions (Figure 9).

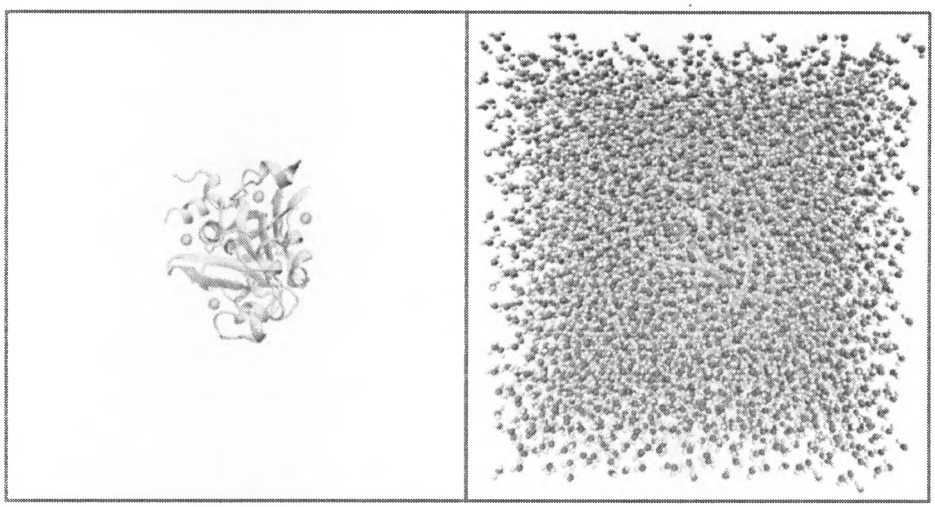


Figure 9. Left panel: hAAG with the addition of Na^+ counterions. Right panel: protein hydration inTIP3P water box.

The last preparative step is hydration of the system by the addition of a TIP3P water box (Figure 9). A water box is a specific set size box which contains water molecules. There are various types of water boxes that may be chosen. Not only can a water box come in different shapes and sizes, it can also vary in the way that the individual water molecules are represented or modeled. For this thesis, the TIP3P water model was chosen because it best coincides with the AMBER program used, and it is also well known for better performance in calculating specific heats. This is especially important to ensure that the

water 'behaves' during simulation the way it would under standard lab conditions.[45] Other water models approximate water with different points.

The TIP3P model approximates water as 3 points, one point for each atom. There are also 4-point and 5-point water models to account for the e lone pairs on oxygen. They are more computational expensive and are not necessary for work done in this thesis (Figure 10). T TIP3P is still considered a highly accurate water model and works well with large protein systems.

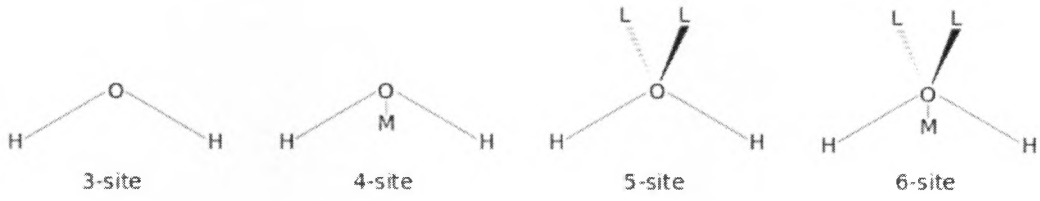


Figure 10. Four different water models commonly used in MD simulations.[25]

2. Minimization, Heating and Equilibration of the System

Minimization was briefly discussed in the theory section of this thesis. To further elaborate, minimization was performed as follows for all three simulations. The first minimization was performed with restraints on enzyme. Restraints are put on the enzyme so that the solvent (water) may have energy minimized first. Once this minimization is done, a minimization is performed for the whole system (without any restraints).

Following minimization is a brief simulation which brings the temperature of the environment up to 310 K from 0 K. The temperature of 310 K is chosen because that is

the average temperature of the human body. Since the heating is also considered as a simulation, parameters such as time-step are also set. For heating, a time-step of 2 fs is also chosen. The volume is held constant while the pressure is not controlled. This allows the temperature to slowly increase without worrying about a negative impact on equilibration. Immediately after heating the system to 310 K, equilibration is run with the pressure and temperature constant and restraints on the enzyme. The goal with equilibration is to allow the density of the water box to reach the normal density of $1.01 g/cm^3$ as initial water packing results in a much lower (0.65-0.7 g/cm^3) density for the box. A second equilibration is run without any restraints.

3. Gathering Production Data

A total of 1 μs of data was gathered over ten 100 ns runs for each of the three simulations. With a time-step of 2 femtoseconds set, this means each of the ten runs done for each simulation contained 50,000,000 snapshots of atomic movement. That brings a total of 500,000,000 computational snapshots for each individual atom throughout the course of the 1 μs data accumulation. With this many snapshots or movements to keep track of, it is highly recommended to save *rstrt* files. What these "restart" files do is create a save point with information regarding current positions. Save points are important in case the system crashes or there is a power outage which causes the simulation to stop. Due to the large number of steps to gather information for, the restart file was saved every 12.5 ns. Depending on the size of the structure being simulated, it

can take 3-6 weeks to gather this data.

A constant temperature (310 K) and pressure (1 bar) were set for the production runs (NTP ensemble). Langevin dynamics were set and act as a thermostat which allows for minor fluctuations due to larger collisions during simulation without drastically effecting the overall simulation.[46]

4. Methods of Data Analysis

Multiple methods of analysis were executed using AMBER tool CPPTRAJ. CPPTRAJ is an updated version of the AMBER tool PTRAJ, which stands for "process trajectory". CPPTRAJ comes with more analysis tools and is primarily written in C/C++.[47] Development of different codes and computational programs are not within the scope of this thesis and therefore won't be discussed any further. All of the following methods of analysis were executed using CPPTRAJ: hydrogen-bond calculations, root-mean-square deviation of distance for all residues, the β-factor for each residue and changes in distance between specific residues to check for stabilizing points. For all hbond analysis, parameters were set to output any hbonds within 3 Å and an angle of 90° at the donor-acceptor interface. RMS (root-mean-square) deviation (or RMSd) measures the deviation of a target set of coordinates to a reference set of coordinates. It is powerful measure to access the dynamic behavior of the system. where δ is the distance between N pairs of equivalent atoms (usually Cα and sometimes C, N, O, C_b). RMSD takes into account how much each atom moves, based on its mass (m_i). The 'heavier' atoms affect structure and

dynamics more. Multiple RMSd plots were obtained for different sets of residues within the system. This is especially handy because it allows the user to observe at what time major deviations occur. The last two calculations are done with AMBER Tools, *bfactor* and *distance.* These calculations are used to further understand which parts of the protein or system are responsible for the large fluctuations in the RMSd. The *bfactor* command plots atomic fluctuations for individual residues. The calculated bfactors indicate the degree of motions for each residue in protein structure. While RMSd data describes the mobility of protein as a whole or over a given range, it may not provide enough information to understand how residue movement affects the protein's mechanism. In addition to *bfactor*, various *distances* can be set to plot the changes in angstroms over the course of the simulation for two sets of 'masks'. In AMBER jargon, a mask is simply the name of one of the atoms/residues being focused on. For various *distance* plots, the two masks are simply distances between two residues. For one of the plots, a mask was set for a range of residues. For that specific type of distance plot, CPPTRAJ uses the range as a collective geometrically, and plots the distance between the geometric center of the range and the other residue. A recap of the overall methods is depicted in Figure 11.

Results & Discussion

Human alkyladenine glycosylase is part of a very important defense system that the human body uses to correct natural mistakes in DNA coding and protects against mutations from exposure to environmental carcinogens. It is already understood that hAAG's defense works through use of BER by glycosidic bond cleavage. The purpose of this thesis has been not only to further investigate the catalytic mechanism for this enzyme, but also to elucidate the conformation of unbound hAAG. The conformational data of the unbound form, together with current data on the hAAG-DNA complex, will be essential for the understanding how this enzyme searches for the sites of DNA damage. Analyzing both bound and unbound conformations should provide valuable structural details for conformational changes needed for transition between free protein and protein complexed to DNA. Figure 12 shows the hAAG-DNA complex used a starting point in our simulations. In this conformation, the catalytic residue Y162 (probe) intercalates into the double helix causing the DNA lesion to be flipped into the glycosylase active site. A water molecule is appropriately positioned to be activated by D238 and act as a nucleophile to catalyze hydrolysis of the *N*-glycosyl bond to release the damaged base. hAAG's active site is lined with hydrophobic, electron-rich aromatic amino acid residues that effectively bind and stabilize positively charged alkylated bases through π-π stacking interactions. The positively charged alkylated bases serve as good leaving groups during nucleophilic elimination.

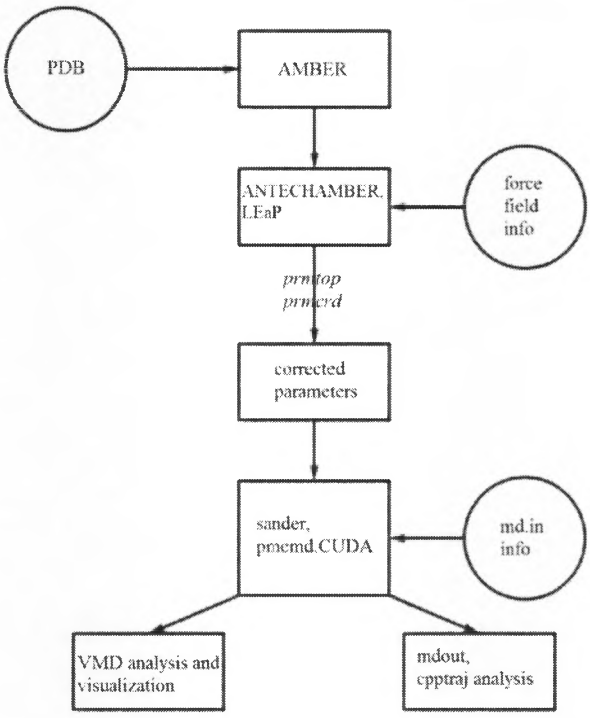


Figure 11. Flow chart of the computational procedure followed for MD simulations of hAAG and hAAG-DNA. complex using AMBER

The detailed analysis of the AAG-DNA complex allows us to identify two protein motifs, besides the active site, which make close contact with lesion-containing DNA. We propose that hAAG-DNA interactions are stabilized by 2 loops: G263-P274 and P130-G148. These loops provide favorable binding interface between the protein and DNA allowing the formation of the high affinity complex.

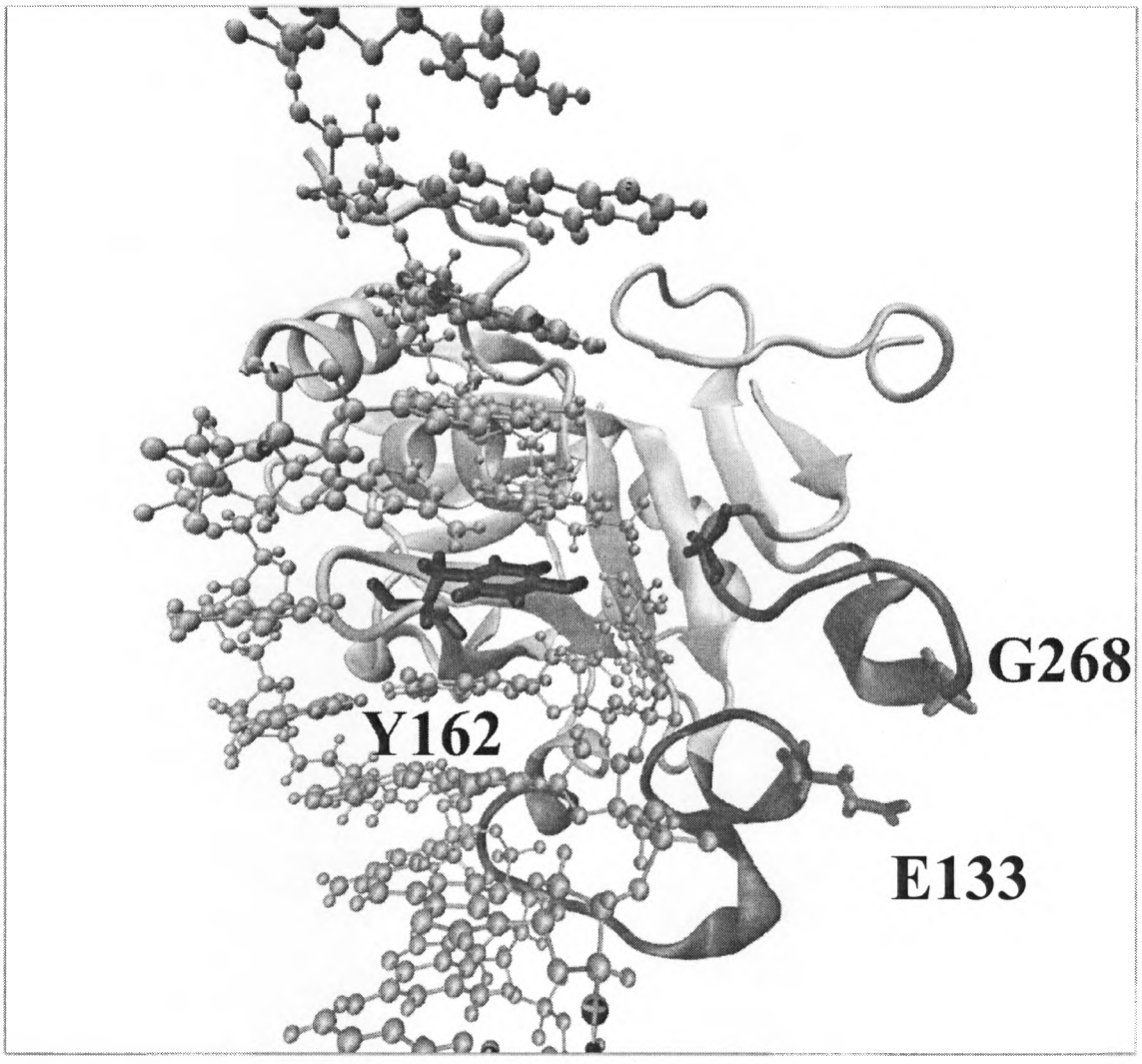


Figure 12. The bound conformation of hAAG demonstrated with hAAG-DNA(eA) complex at 0 ns. The probe, Y162, is shown in black while two residues from G263-P274 and P130-G148 loops are show in dark gray. The protein and DNA are shown in light

Upon completion of the production run for the bound and unbound conformations, RMSd plots were processed. (Appendix item 1,3). The RMSd profiles showed comparable deviations for the unbound AAG and AAG complexed to DNA. This finding indicates that there no large conformational change in the overall structure for both systems. Further analysis of the structure with VMD led to discovering mobility in a G263-P374 loop adjacent to the active site in the bound conformation. For the remainder of this paper, we refer to this loop as "loop". In Figure 13, the G263-P274 loop is shown in darker shade on the right of the protein structure.

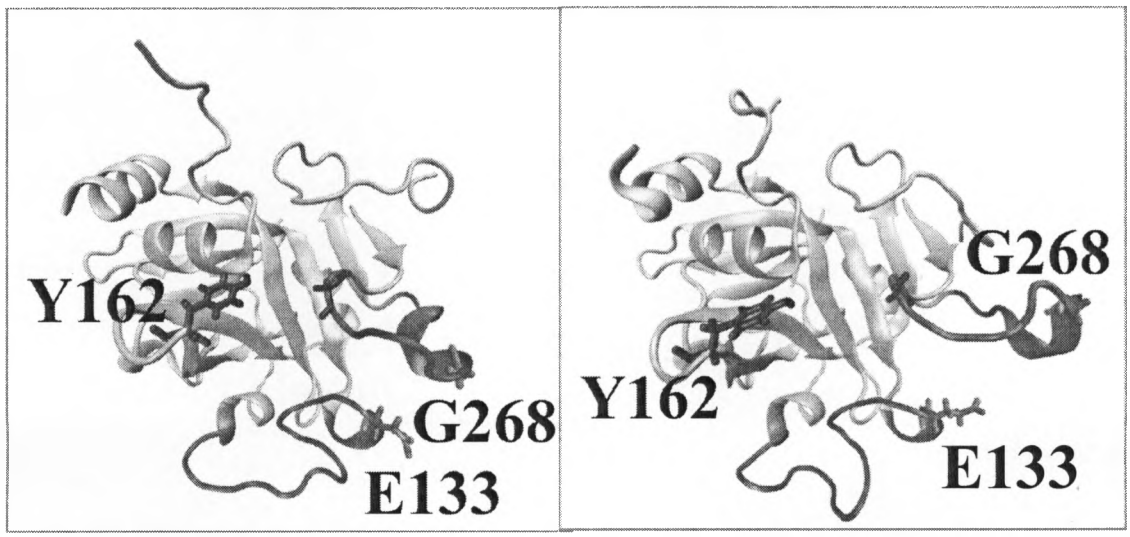


Figure 13. Two snapshots of the unbound (wt) conformation of hAAG: (left) a 'closed' conformation at 0 ns and (right) a more 'open' conformation at 1000 ns.

In Figure 14, the bound conformation is shown with (14a) and without (14b) the DNA pictured. They are the same structurally, but the DNA was simply removed from the image to provide a better picture to compare with the unbound form of hAAG. The bound

and unbound images appear to be quite similar. While it was initially thought that the unbound structure would point to more loop mobility, the bound structure actually had slightly more loop mobility. This is supported by the RMSD and b-factor data mentioned previously, where there was a slight increase in distance between the loop and the probe with the DNA present, as well as increased mobility in protein residues.

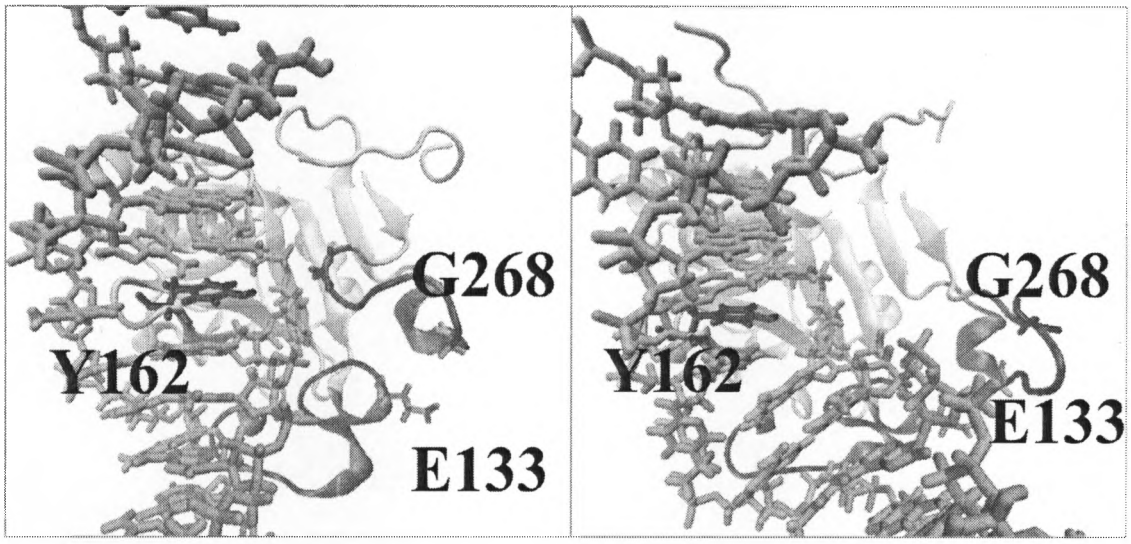


Figure 14a. Two snapshots of the bound conformation of hAAG: (left) a 'closed' conformation at 0 ns and (right) a more 'open' conformation at 1000 ns.

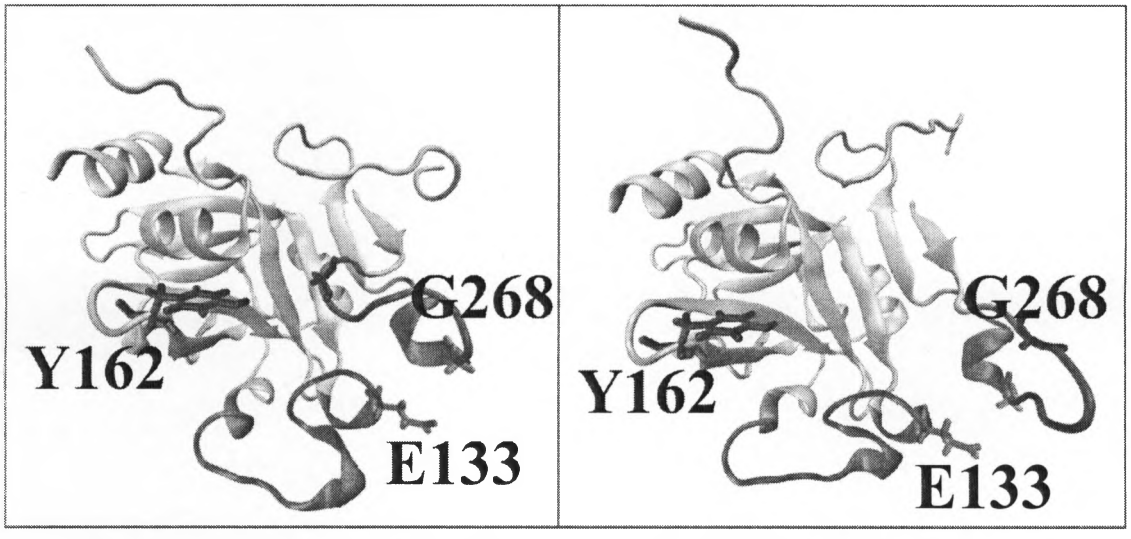


Figure 14b. Two snapshots of the bound conformation of hAAG: (left) a 'closed' conformation at 0 ns and (right) a more 'open' conformation at 1000 ns. The DNA portion of the bound complex is not shown so that protein changes may be more easily observable.

The next step in our analysis was to calculate β-factors for both systems. Both unbound and bound AAG simulations reveal high flexibility of the loop region (Appendix 7, 9). The analysis of the unbound conformation showed much large β-factors (particularly for residue G268) as compared to the bound structure (Appendix 8, 10). This indicates increased mobility of G263-P274 loop prior to formation of the AAG-DNA complex. Another sequence of residues, P130-G148, is referred to as the 'bottom loop'. This bottom loop is also adjacent to the active site and plays a role in holding the DNA in place. There were also noticeable ranges of motion for the bottom loop. RMSd data was obtained for the loop and bottom loop. The loop had a maximum RMSd value of 10.09 angstroms and 8.41 in the bound and unbound conformations, respectively (Appendix 2, 5). The bottom loop had a maximum RMSd value of 3.26 and 4.90 angstroms in the bound and unbound conformations, respectively. One common trend seen for all RMSd

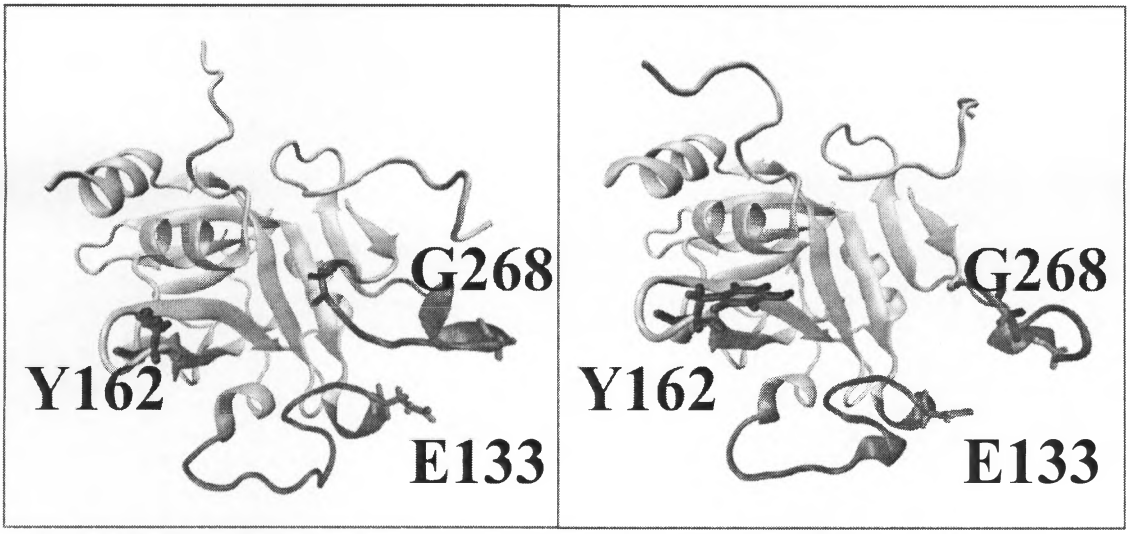


Figure 15. Two snapshots taken at 200 ns simulation time: (left) unbound conformation and (right) bound conformation. The DNA portion of the bound complex is not shown so that protein changes may be more easily observable.

plots was a spike around 200 ns. Figure 15 shows a snapshot of both conformations at 200 ns simulation time.

Generally, there are two perspectives on the overall behavior of hAAG in its unbound conformation and how it changes upon complex formation with DNA. The first idea is that there is a sliding mechanism by which hAAG scans along the DNA looking for or waiting to encounter a lesion which acts as a substrate for its active site.[48] The second perspective is one in which the DNA repair enzyme hops or clamps along the DNA, waiting to encounter a lesion to repair. Thymine DNA glycosylase, an enzyme from the same family as hAAG, was studied by Buechner et al. with atomic force microscopy. The main goal of their study was to investigate whether a scanning or hopping mechanism was more probable. They did employ computational methods; however, they primarily used other non-computational methods, and concluded that there were some characteristics of 'hopping' behavior for the enzyme along DNA, but mainly a scanning behavior. A scanning mechanism appears to be a more efficient behavior. Clamping or hopping would lead to a much more inefficient rate of BER.[49]

Lastly, Jeremy Setser and his associates proposed hAAG macromolecular behavior cannot be classified as either "scanning" or "hopping". Rather, they support hAAG utilizing both behaviors—passively scanning the DNA, and actively further clamping down upon the lesion site on DNA surface.[50] They used ESP mapping and some computational methods primarily known as "CNS". CNS, or crystallography and

NMR system is a software suite for macromolecular structure determination, which has similar components to AMBER.[43] Both of these case studies reflect information also seen throughout the analysis of hAAG in bound and unbound conformations throughout this thesis. Our initial hypothesis was that the unbound hAAG would stay in the open conformation when both loops extends toward the incoming DNA. The nonspecific DNA binding interactions will promote loop closing around the DNA with the modified base interacting with the enzyme active site. This tight binding together with existence of the well-defined open conformation for unbound hAAG should support clamping (hopping) mechanism. In Such mechanism protein probes for the lesion by microscopic dissociation and association events that allow the protein to sample both DNA backbones .. Instead, there were no distinguishable differences between the behavior of the loop and bottom loop within the bound and unbound conformations. Interesting that both loops revealed high structural flexibility in the AAG-DNA complex. Moreover, the slightly higher mobility was observed for the G263-P724 loop adjacent to the active site, than P130-G148 loop. This noticeably flexibility of both loops when DNA is present, may indicate that the binding of the DNA is not as tight as expected. A more rigid loop structure would not be favorable in the case of a clamping mechanism because then the loop would not be free to open and close as needed. A more rigid loop structure would be indicative of a scanning mechanism. This processive scanning will allow AAG to remove multiple base lesions from in a single binding encounter. These arguments agreed with the studies done by Jeremy Setser and Buechner et al.

Two additional and very important methods of analysis performed were *hbond* analysis and *distance* analysis. The aim of using the *hbond* function of CPPTRAJ was to highlight any changes in h-bonds when going from bound to unbound structure. It was proposed that there would be h-bonds present in the bound conformation that would be missing once the DNA was removed. Table 1 below shows the h-bonds found between protein and loop, protein and DNA, loop and DNA, and between R197 and protein/nucleic components. The reason R197 is specified is because R197 was one h-bond found in the protein-DNA h-bond analysis.

Table 1. H-bond contacts detected by CPPTRAJ within 90° and 3Å.

	bound		unbound	
protein-loop	V262	G263	V262	G263
	L275	P274	L275	P274
protein-DNA	R197	DG10	-	-
loop-dna	-	-	-	-
R197-all	R197	S198	R197	S198
	R197	E213	-	-
	R197	DG10	-	-
	R197	L196	R197	L196
	R197	Q223	-	-
	R197	M193	-	-

The results shown in Table 1 came as a surprise. With such a difference in structure, it was expected there would be numerous h-bond contacts missing in the unbound conformation of hAAG. Instead, the major changes detected when shifting from bound to

unbound conformation all involved a single residue, R197. It appears that R197 acts as an anchor for the DNA while BER is executed. The absence of the multiple h-bond contacts in the interface between the protein and DNA also supports mechanism when bound AAG diffuses down the DNA in search of the damages sites. With additional h-bond contacts, it would be much harder for hAAG to execute processive searching mechanism along the DNA.[51]

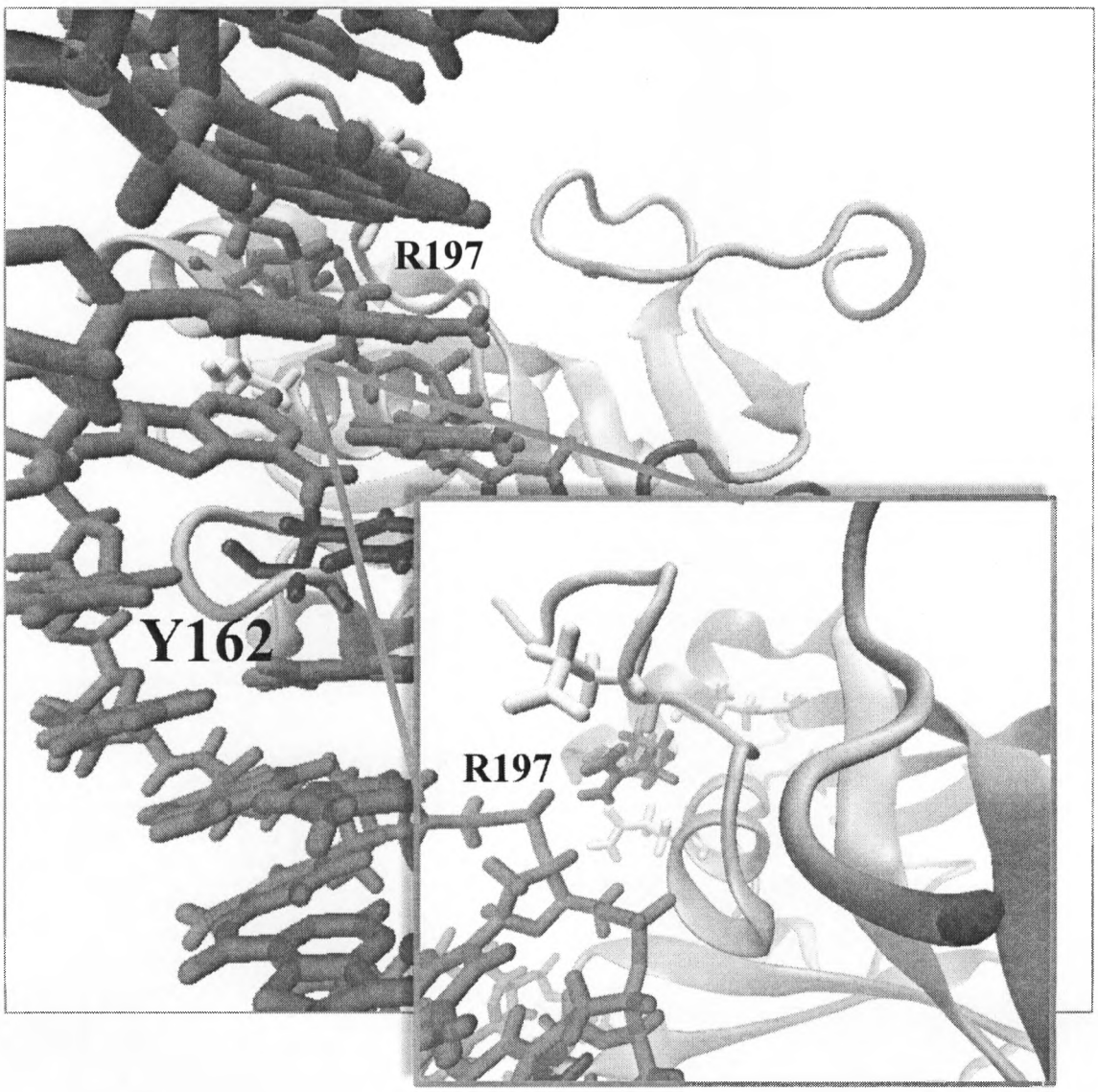


Figure 16. A snapshot of hydrogen contacts found between R197 and residues DG10, E213, Q223 and M193. These contacts are found at the interface between DNA and hAAG, above and away from the protein active site.

Conclusion

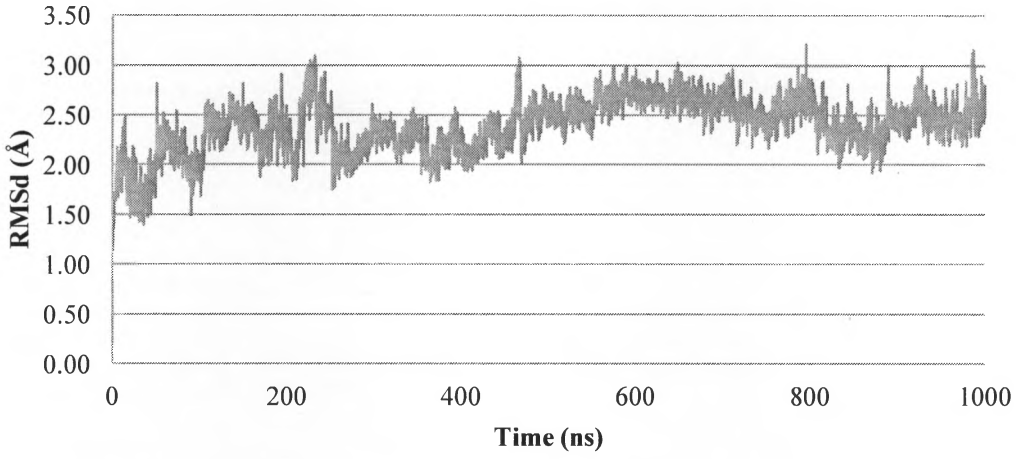
The hAAG enzyme plays a primary role in repairing alkylation DNA damage which have been linked with many cancers, including liver and colorectal. Moreover, given that many cancer chemotherapy treatments intentionally produce alkylation DNA damage to kill rapidly growing cells, hAAG has been investigated as a potential biomarker for therapeutic response. The scope of this work employed molecular dynamics simulation to provide structural insights the into mechanism of how the human AAG searches for rare lesion substrates amongst the vast excess of normal undamaged bases. An important part of such work was to perform structural comparison between unbound and bound forms of the hAAG. However, currently there is no known structure for the wild-type, or unbound conformation, of hAAG at an atomic resolution. This is a vital piece of information which is missing from the vast amount of knowledge on how hAAG functions.

Under this project, using computationally approaches, we successfully perform explicit solvent simulation of the wild-type hAAG without bound DNA. The conformational dynamics of the bound and unbound hAAG clearly indicate that the formation of the protein complex does not lead to a tight binding between protein and DNA. Based on the RMSd and β-factors analysis, the two loops adjacent to the enzyme active site revealed significant conformational flexibility in the bound conformation. This flexibility was comparable or even lower for the corresponding structural motifs in the unbound simulations. The observed relaxed protein conformation around the DNA

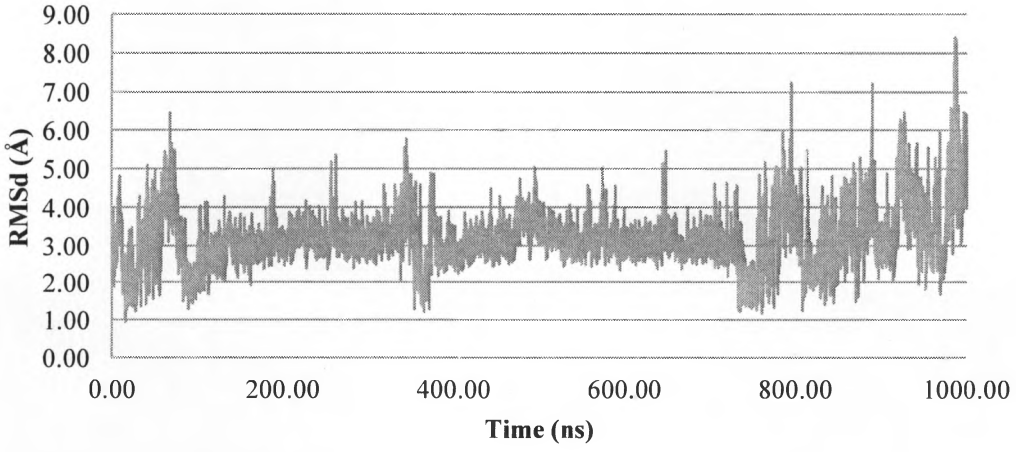
should support processive scanning mechanism in which bound protein slides down the DNA structure in search of damaged sites. Such "loose" interaction was also supported by the observation of the limited number of hydrogen bond contacts in the DNA-protein interface. In the future, the characterization of hAAG-DNA interactions can be improved by calculating the binding free energies between the protein and substrate or extending the MD simulations to the higher time scales. Since the success of such calculation will rely on initial explicit solvent calculation (unbound and bound hAAG) this project provides an excellent foundation for future work.

The hope is that enough information will eventually be gathered by studying the general mechanism and interactions found during complex formation between hAAG and DNA. Adequate understanding of how hAAG operates before, during, and after DNA complex formation may lead to future cancer research down the path of pharmaceutical development.

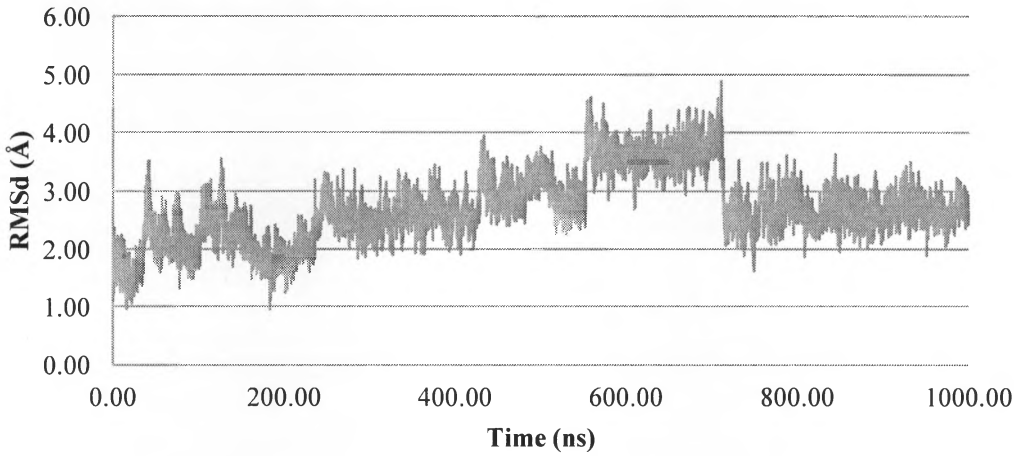
APPENDICES



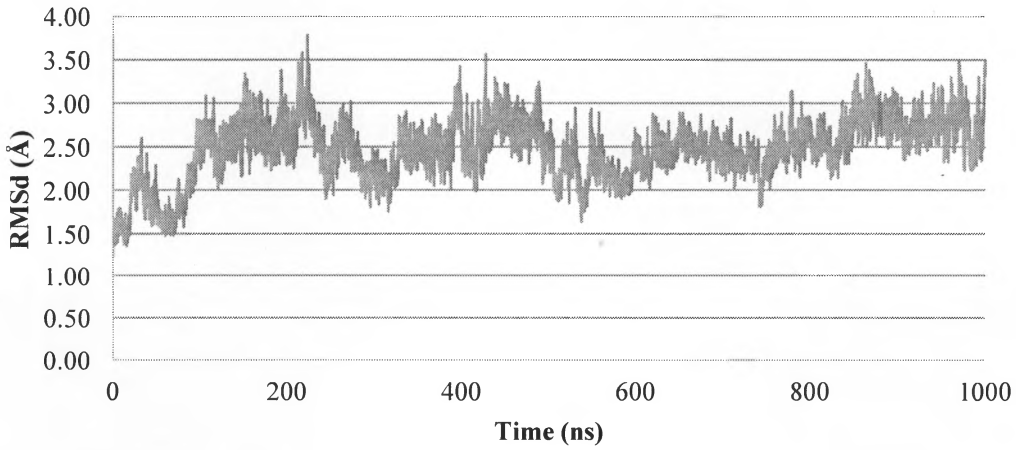
1. RMSd of hAAG. Root-mean-square deviation of the positions of individual residues in unbound (wt) hAAG over the course of 1000 ns.



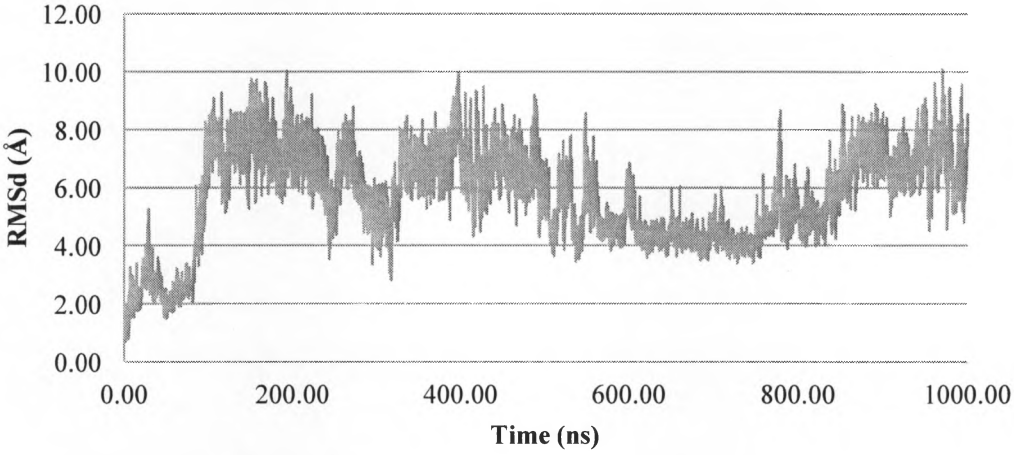
2. RMSd of loop in hAAG. Root-mean-square deviation of the positions of residues G263-P274 in unbound (wt) hAAG over the course of 1000 ns.



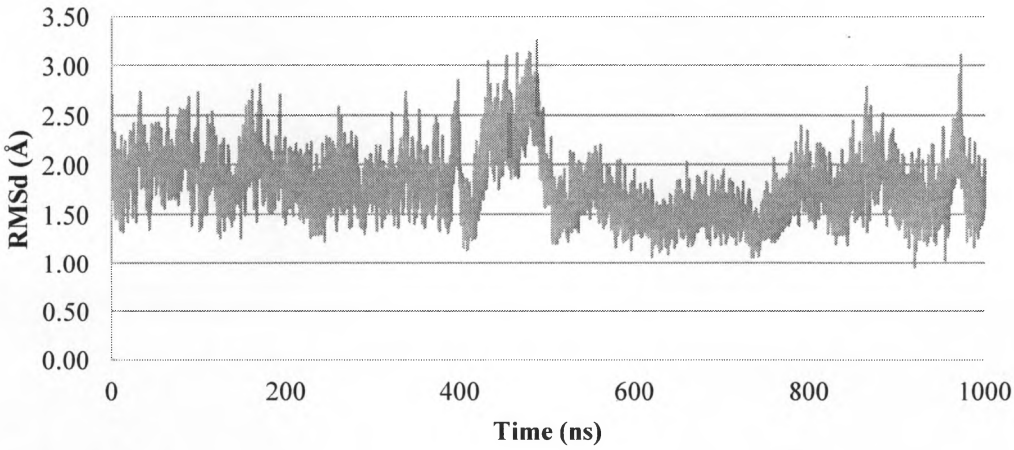
3. RMSd of bottom loop in hAAG. Root-mean-square deviation of the positions of residues P130-G148 in unbound (wt) hAAG over the course of 1000 ns.



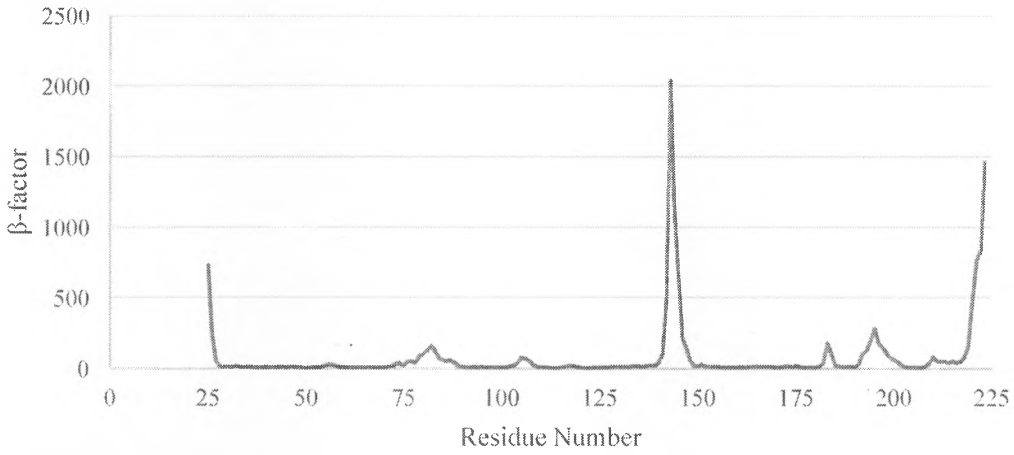
4. RMSd of hAAG-DNA(εA). Root-mean-square deviation for all residues of DNA-bound hAAG with an εA lesion in the active site over the course of 1000 ns.



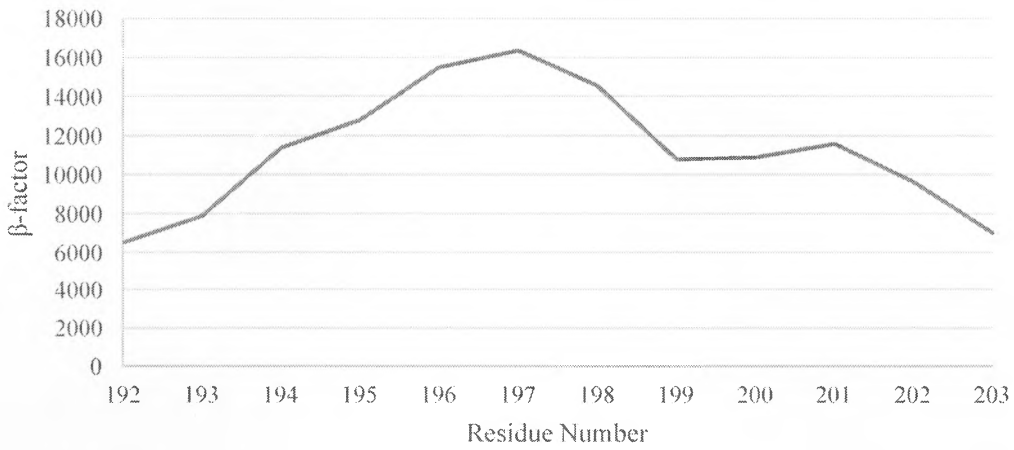
5. RMSd of loop in hAAG-DNA(εA). Root-mean-square deviation of the positions of residues G263-P274 in DNA-bound hAAG with an εA lesion in the active site over the course of 1000 ns.



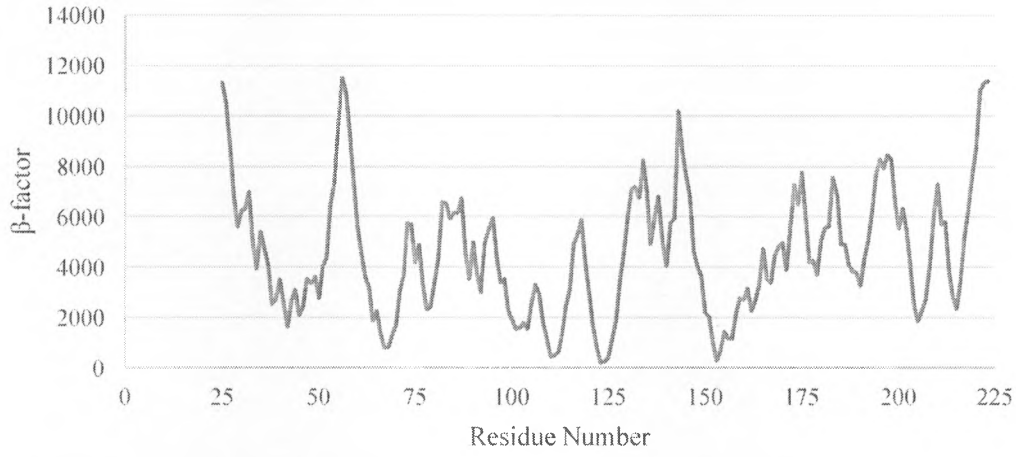
6. RMSd of bottom loop in hAAG-DNA(εA). Root-mean-square deviation of the positions of residues P130-G148 in DNA-bound hAAG with an εA lesion in the active site over the course of 1000 ns.



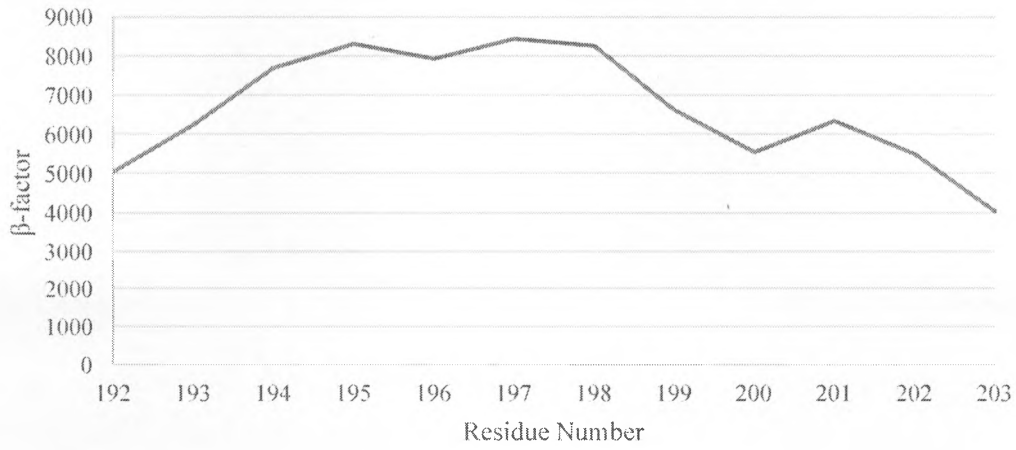
7. β-factor for hAAG. B-factor for all residues, H82-Q294, in unbound (wt) hAAG.



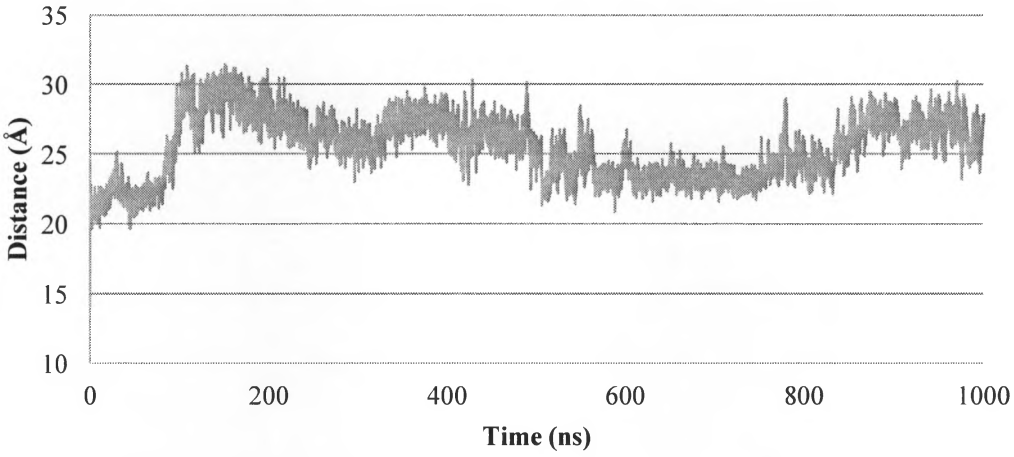
8. β-factor for the loop in hAAG. B-factor for loop residues G263-P274, in unbound (wt) hAAG.



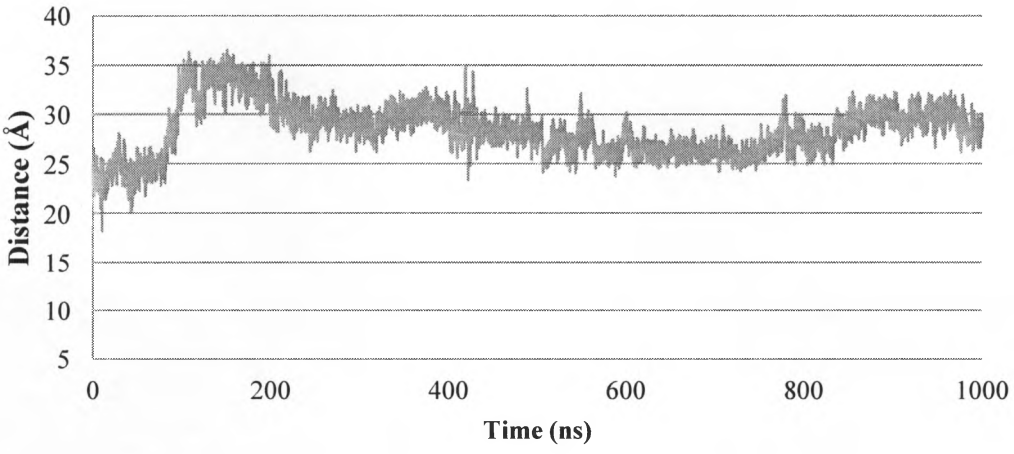
9. β-factor for hAAG-DNA(εA). B-factor for all protein residues, H82-Q294, in DNA-bound hAAG with an εA lesion in the active site.



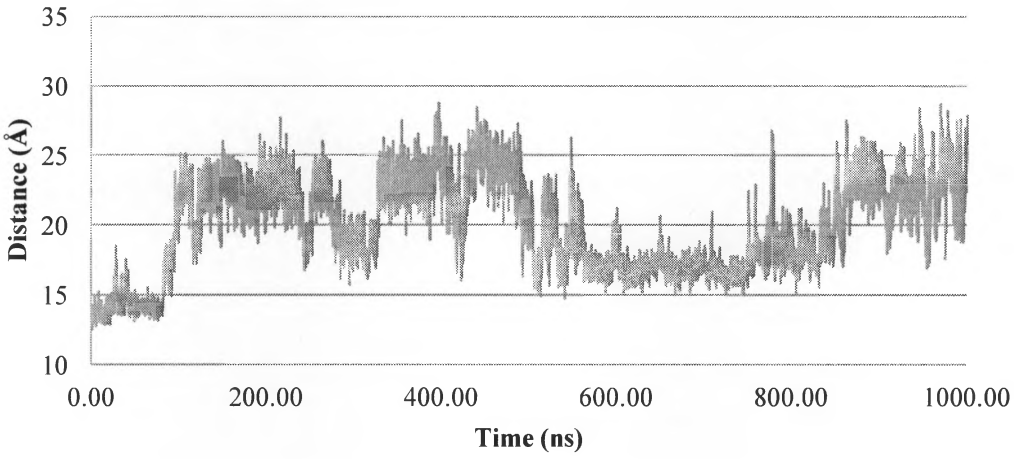
10. β-factor for the loop in hAAG-DNA(εA). B-factor for loop residues G263-P274, in DNA-bound hAAG with an εA lesion in the active site.



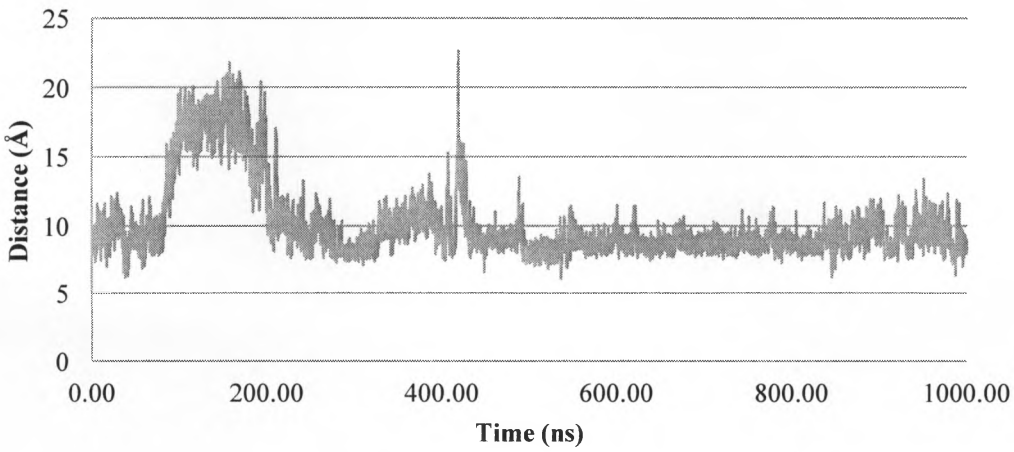
11. Variation in loop-probe distance for hAAG-DNA(εA). The changes in distance between loop residues G263-P274 and the probe, Y162, in DNA-bound hAAG with an εA lesion in the active site.



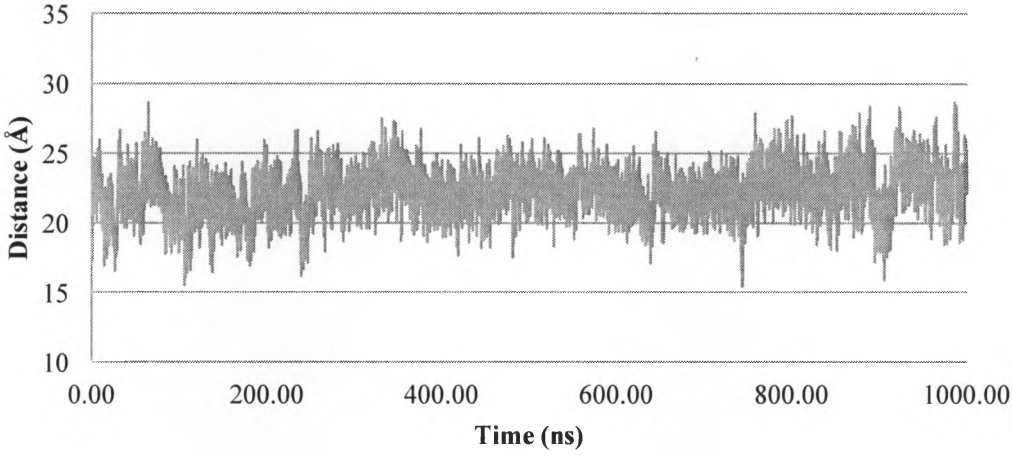
12. Variation in G268-probe distance for hAAG-DNA(εA). The changes in distance between loop residue G268 and the probe, Y162, in DNA-bound hAAG with an εA lesion in the active site.



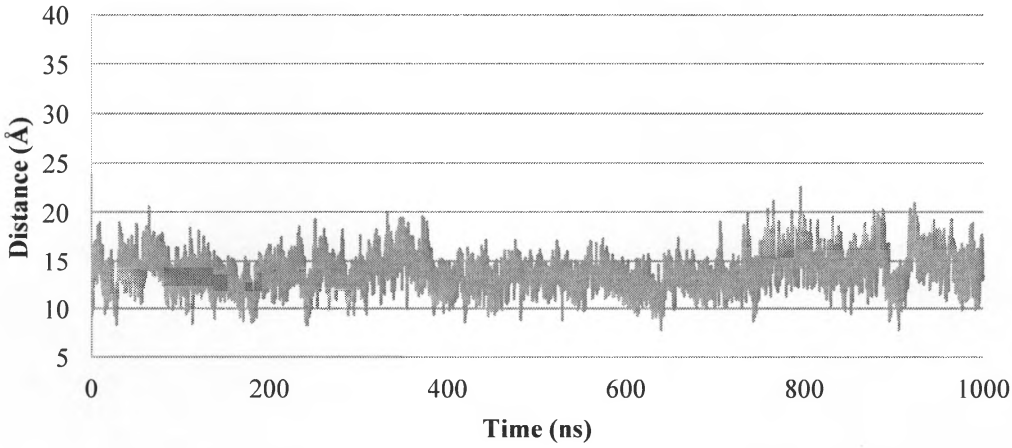
13. Variation in G263-probe distance for hAAG-DNA(εA). The changes in distance between loop residues G263 and the probe, Y162, in DNA-bound hAAG with an εA lesion in the active site.



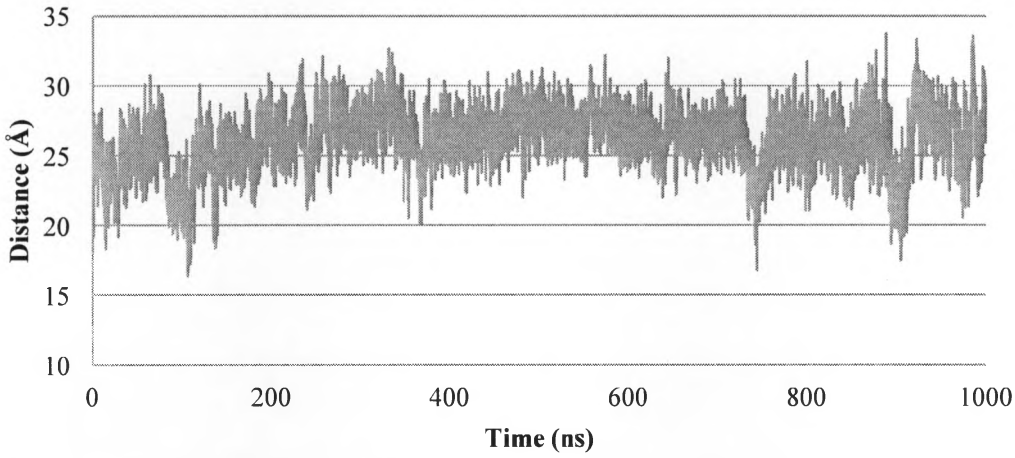
14. Variation in G268-E133 distance for hAAG-DNA(εA). The changes in distance between residues 197 and 76 in DNA-bound hAAG with an εA lesion in the active



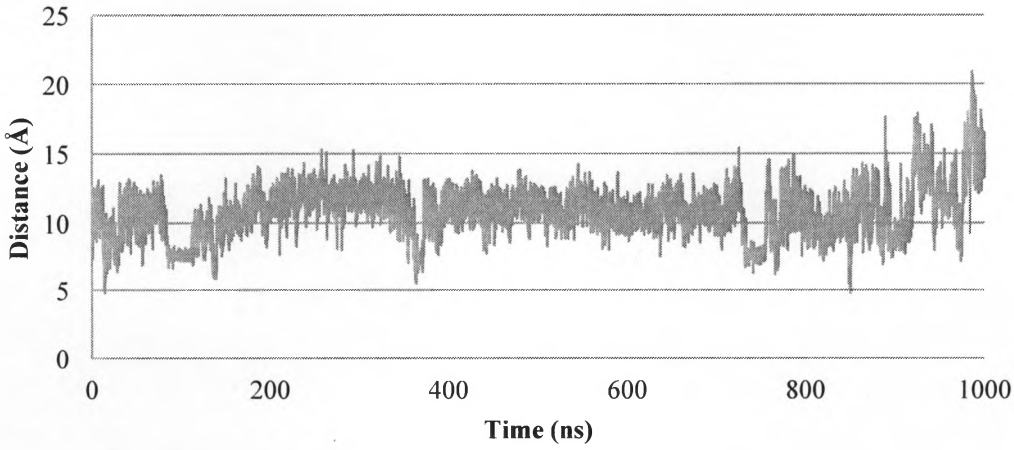
15. Variation in loop-probe distance for hAAG. The changes in distance between loop residues G263-P274 and the probe, Y162, in unbound (wt) hAAG.



16. Variation in G263-probe distance for hAAG. The changes in distance between loop residue G263 and the probe, Y162, in unbound (wt) hAAG.



17. Variation in G263-probe distance for hAAG. The changes in distance between loop residue G263 and the probe, Y162, in unbound (wt) hAAG.



18. Variation in G268-E133 distance for hAAG. The changes in distance between loop residue G268 and residue E133 in unbound (wt) hAAG.

(1) World Health Organization. *Cancer Fact Sheet*; 2017.
(2) Laib, R. J. *IARC* **1986**, *70*, 101–108.
(3) Rutledge, L. R.; Wetmore, S. D. *J. Am. Chem. Soc.* **2011**, *133* (40), 16258–16269.
(4) *Smokeless tobacco and some tobacco-specific N-nitrosamines*; IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, International Agency for Research on Cancer, Eds.; IARC monographs on the evaluation of carcinogenic risks to humans; World Health Organization, International Agency for Research on Cancer ; distributed by WHO Press: Lyon, France : Geneva, 2007.
(5) *Some monomers, plastics and synthetic elastomers, and acrolein: this publication represents the views and expert opinions of an IARC Working Group on the Evaluation of the Carcinogenic Risk of Chemicals to Humans which met in Lyon, 7 - 13 February 1978*; International Agency for Research on Cancer, Ed.; IARC Monographs on the evaluation of the carcinogenic risk of chemicals to humans; IARC: Lyon, 1979.
(6) Baba, M.; Yamamoto, R.; Iishi, H.; Masaharu, T. *Int.J.Cancer* **1997**, *72*, 815–820.
(7) Drablos, F.; Emadoldin, F.; Per Arne, A. *DNA Repair* **2004**, *3* (11), 1389–1407.
(8) Briggs, D. *Br. Med. Bull.* **2003**, *68* (1), 1–24.
(9) Wang, P.; Guliaev, A. B.; Hang, B. *Toxicol. Lett.* **2006**, *166* (3), 237–247.
(10) Barbin, A. *Mutat. Res* **2000**, *462*, 55–69.
(11) Branze, D.; Foiani, M. *Nat. Rev. Mol. Cell Biol.* **2008**, *9*, 297–308.
(12) Abner, C. W. *J. Biol. Chem.* **2001**, *276* (16), 13379–13387.
(13) Friedman, J. I.; Stivers, J. T. *Biochemistry (Mosc.)* **2010**, *49* (24), 4957–4967.
(14) Hollis, T.; Lau, A.; Ellenberger, T. *Mutat. Res. Repair* **2000**, *460* (3), 201–210.
(15) Wolfe, A. E. Kinetic Mechanism for Binding and Flipping of Damaged Bases By Alkyladenine DNA Glycosylase, The University of Michigan, 2010.
(16) Vallur, A. C. *J. Biol. Chem.* **2002**, *277* (35), 31673–31678.
(17) Vallur, A. C.; Maher, R. L.; Bloom, L. B. *DNA Repair* **2005**, *4* (10), 1088–1098.
(18) Brooks, S. C.; Adhikary, S.; Rubinson, E. H.; Eichman, B. F. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **2013**, *1834* (1), 247–271.
(19) Clancy, S. *DNA Repair Mech. Detect Repair Damaged DNA What Happens They Fail* **2008**, 10–4.
(20) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21* (6), 1087.
(21) Levitt, M.; Warshel, A. *Nature* **1975**, *253* (5494), 694–698.
(22) Nobel Media AB 2014. *The Nobel Prize in Chemistry 2013 - Press Release*; 2013.
(23) Tuckerman, M. E.; Martyna, G. J. *J. Phys. Chem. B* **2000**, *104* (2), 159–178.
(24) Weiner, P. K.; Kollman, P. A. *J. Comput. Chem.* **1981**, *2* (3), 287–303.
(25) Leach, A. R. *Molecular Modelling: Principle and Applications*; Prentice Hall: Great Britain, 2001.
(26) Levitt, M.; Hirshberg, M.; Sharon, R.; Daggett, V. *Comput. Phys. Commun.* **1995**, *91* (1–3), 215–231.
(27) Herrebout, W. A. *J. Phys. Chem.* **1995**, *99* (2), 578–585.

(28) Verlet, L. *Phys. Rev.* **1967**, *159* (1), 98.
(29) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. *Comput. Phys. Commun.* **1995**, *91* (1–3), 1–41.
(30) Beeman, D. *Journal of computation Physics* **1976**, *20.2*, 130–139.
(31) Dong, F.; Olsen, B.; Baker, N. A. *Methods in cell biology* **2008**, *84*, 843–870.
(32) Masunov, A.; Lazaridis, T. *J. Am. Chem. Soc.* **2003**, *125* (7), 1722–1730.
(33) Bureau, H. R.; Merz Jr, D. R.; Hershkovits, E.; Quirk, S.; Hernandez, R. *PloS One* **2015**, *10* (5), e0127034.
(34) Becker, O. M.; Mackerel, Jr., A. D.; Roux, B.; Watanabe, M. *Computational Biochemistry and Biophysics*; Marcel Dekker, Inc.: New York, 2001.
(35) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T.; others. *J Am Chem Soc* **1990**, *112* (16), 6127–6129.
(36) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23* (14), 1297–1304.
(37) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4* (2), 187–217.
(38) Zwier, M. C.; Chong, L. T. *Curr. Opin. Pharmacol.* **2010**, *10* (6), 745–752.
(39) Jo, J. C.; Kim, B. C. *Bull. Korean Chem. Soc.* **2000**, *21* (4), 419–424.
(40) Basma, M.; Sundara, S.; Çalgan, D.; Vernali, T.; Woods, R. J. *J. Comput. Chem.* **2001**, *22* (11), 1125–1137.
(41) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33–38.
(42) Lau, A. Y.; Wyatt, M. D.; Glassner, B. J.; Samson, L. D.; Ellenberger, T. *Proc. Natl. Acad. Sci.* **2000**, *97* (25), 13573–13578.
(43) Brünger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J.-S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; others. *Acta Crystallogr. D Biol. Crystallogr.* **1998**, *54* (5), 905–921.
(44) Drabik, P.; Liwo, A.; Czaplewski, C.; Ciarkowski, J. *Protein Eng.* **2001**, *14* (10), 747–752.
(45) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79* (2), 926–935.
(46) Schlick, T. *Molecular modeling and simulation: an interdisciplinary guide*, 2nd ed.; Interdisciplinary applied mathematics; Springer: New York, 2010.
(47) Roe, D. R.; Cheatham, T. E. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095.
(48) Hedglin, M.; O'Brien, P. J. *Biochemistry (Mosc.)* **2008**, *47* (44), 11434–11445.
(49) Buechner, C. N.; Maiti, A.; Drohat, A. C.; Tessmer, I. *Nucleic Acids Res.* **2015**, *43* (5), 2716–2729.
(50) Setser, J. W.; Lingaraju, G. M.; Davis, C. A.; Samson, L. D.; Drennan, C. L. *Biochemistry (Mosc.)* **2012**, *51* (1), 382–390.
(51) Hubbard, R. E.; Kamran Haider, M. In *Encyclopedia of Life Sciences*; John Wiley & Sons, Ltd, Ed.; John Wiley & Sons, Ltd: Chichester, UK, 2010.