COMPUTATIONAL RATIONALE FOR THE SELECTIVE INHIBITION OF THE HERPES SIMPLEX VIRUS TYPE 1 URACIL-DNA GLYCOSYLASE ENZYME

A dissertation submitted in accordance with the requirements of the

UNIVERSITY OF CAPE TOWN,

in fulfilment of the requirements for the degree of

MASTER of SCIENCE

by

Umraan Hendricks

Supervisor: Professor Kevin J. Naidoo

February 2011

DECLARATION

I declare that COMPUTATIONAL RATIONALE FOR THE SELECTIVE INHIBITION OF THE HERPES SIMPLEX VIRUSTYPE 1 URACIL-DNA GLYCOSYLASE ENZYME is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

Umraan Hendricks

ACKNOWLEDGMENTS

I would like to thank:

1. My supervisor, Associate Professor, Kevin J. Naidoo for his suggestions, support and guidance throughout the project.
2. My fellow group member Ranga Jayakody for his guidance, as well as all the member of the S.C.R.U. group for their recommendations and support.
3. To my wife Rachma Hendricks, a special thank you is required for all the patience you have had over the writing up period and for proof reading my thesis.
4. To my parents for their support, financially and emotionally, and especially to my mother, for her tolerance.
5. Louise Bezuidenhout for all her administrative assistance.

ABSTRACT

The herpes simplex virus uracil-DNA glycosylase (hsvUDG) enzyme is responsible for reactivation of the virus from latency, and efficient viral replication in nerve tissue. The lack of uracil-DNA glycosylase enzyme in human neurons and the continuous deamination of cytosine creates an environment where the presence of viral uracil-DNA glycosylase is a necessity for the proliferation of the virus. 6-(4-Alkylanilino)-uracil inhibitors have been developed that selectively and strongly bind to the hsvUDG enzyme while weakly binding to human uracil-DNA glycosylase (hUDG). In this thesis I will investigate the binding pocket and analyse the nature of binding of the 6-(4-Alkylanilino)-uracil inhibitors in hsvUDG and hUDG to provide a platform for the development of improved inhibitors.

Computational methods were used to analyse the effect of the6-(4-Alkylanilino)-uracil inhibitors on the hsvUDG and hUDG enzymes. Parameters were developed, optimized and validated for the inhibitors using Free Energy Perturbation (FEP) methods and experimental data. These parameters were used to produce 10ns of production molecular dynamics simulation for a selection of inhibitor:hsvUDG and inhibitor:hUDG complexes. To further understand the natural behaviour of the protein, simulations of the DNA:hUDG and DNA:hsvUDG complexes were performed, allowing for the identification of key amino acids. Binding pocket analyses revealed that the binding pocket of the hUDG enzyme is approximately 20Å^3 smaller than that of the hsvUDG enzyme. Two hydrophobic pockets were also discovered in hsvUDG. The primary hydrophobic pocket responsible for the tight binding of the inhibitors was determined not to be present in hUDG. Now that the binding model for these inhibitors has been resolved, it is possible to improve on the 6-(4-Alkylanilino)-uracil inhibitors through further experimental and computational methods.

ABBREVIATIONS

Å	Angstrom
BER	Base Excision Repair
CHARMM	Chemistry at Harvard Macromolecular Mechanics
DNA	Deoxyribose Nucleic Acid
dTTP	Deoxythymidine Triphosphate
dUTP	DeoxyuridineTriphospate
FEP	Free Energy Perturbation
HF	Hartree-Fock
hsvUDG	Herpes Simplex Virus Type 1 Uracil-DNA Glycosylase
hUDG	Human Uracil-DNA Glycosylase
LCAO	Linear Combination of Atomic Orbitals
MD	Molecular Dynamics
MM	Molecular Mechanics
NPT	Isothermal-isobaric Ensemble
ns	Nanosecond
NVE	Microcanonical Ensemble
NVT	Canonical Ensemble
PBC	Periodic Boundary Conditions
ps	Picosecond
QM	Quantum Mechanics

RMSD	Root Mean Squared Deviation
RNA	Ribose Nucleic Acid
SCF	Self Consistent Field
SG	Slow Growth
TI	Thermodynamic Integration

Contents

Chapter 1

Exploring Proteins

1.1 Introducing Enzymatic Proteins

The term enzyme was first used by Friedrich Whilhelm Kuhne in 1878 to describe catalytically active substances that had previously been called ferments.[1] Enzymes are proteins that serve as biological catalysts, that is, they speed up chemical reactions without undergoing any overall chemical change during the reaction. Without enzymes, most metabolic reactions would simply proceed too slowly at normal body temperature to support life. Amino acid residues are the building blocks of enzymes and are connected to each other by peptide bonds to form long polypeptide chains (Figure 1.1).

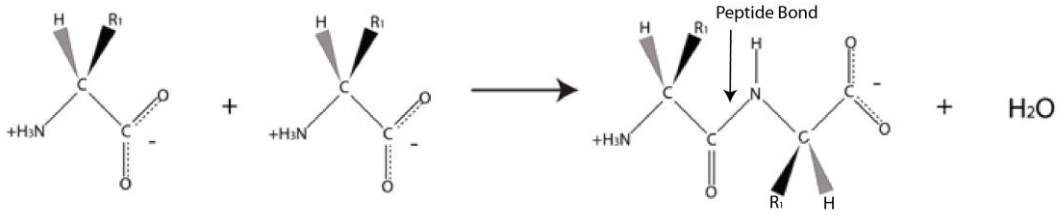
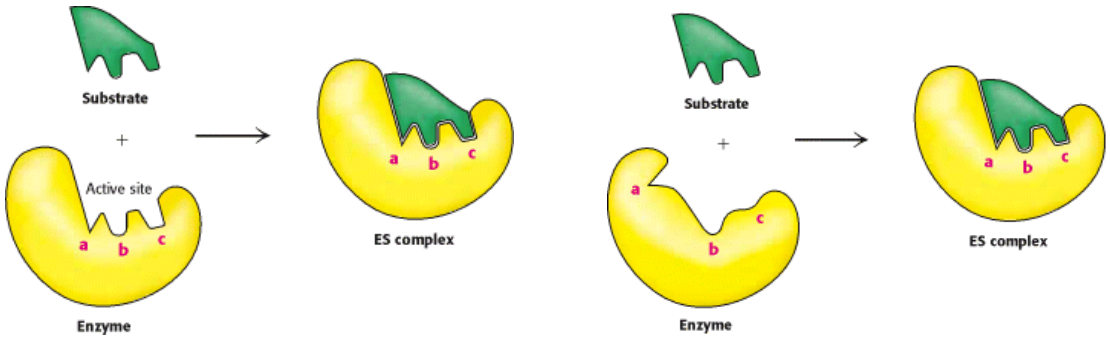


Figure 1.1 Peptide-bond formation. Two amino acids link together with the loss of a water molecule.

Peptide chains fold in highly specific ways that confer 3-dimensional structure to the protein. Enzymes act by attaching to reaction molecules called substrate, as displayed in Figure 1.2 by either the lock and key or induced fit model.[1] Enzymes are highly specific, meaning that each enzyme catalyses only a single reaction, or a very limited class of reactions. The region of the enzyme that binds the substrate is known as the binding pocket or active site. The specific 3-dimensional shape of an enzyme is such that only the substrates it acts upon can "fit" into the binding pocket. The protein residues that directly interact with the substrate are located in the binding pocket and are responsible for the high selectivity of the enzyme.



(A) (B)

Figure 1.2 (A) Lock and key model (B) Induced fit model.[1]

Specific amino acids within the active site are known as catalytic residues. The substrate has to have complementary functional groups that are perfectly orientated and will allow these catalytic residues to interact with them. After catalysing the reaction, the enzyme releases the products of the reaction. The enzyme remains intact in the process and can immediately bind a fresh substrate. Thus, an enzyme molecule can be used over and over again. Enzymes regulate nearly all metabolic activities and are responsible for the building of complex molecules, as well as the breakdown of large molecules into smaller ones, known as anabolic and catabolic processes respectively. Enzymes increase the rate of chemical reactions by lowering the activation energy required for the reaction to proceed in the forward direction.[1]

1.2 Protein Structure

Proteins are linear polymers built of monomer units, known as amino acids. There are a total of 20 naturally occurring amino acids, each with unique molecular properties. Amino acids consist of a carboxylic group, a hydrogen atom, a amino group and a distinctive R group, which are all connected to a central carbon atom commonly referred to as the α carbon. Figure 1.3 shows the chiral L and D isomers of amino acids.

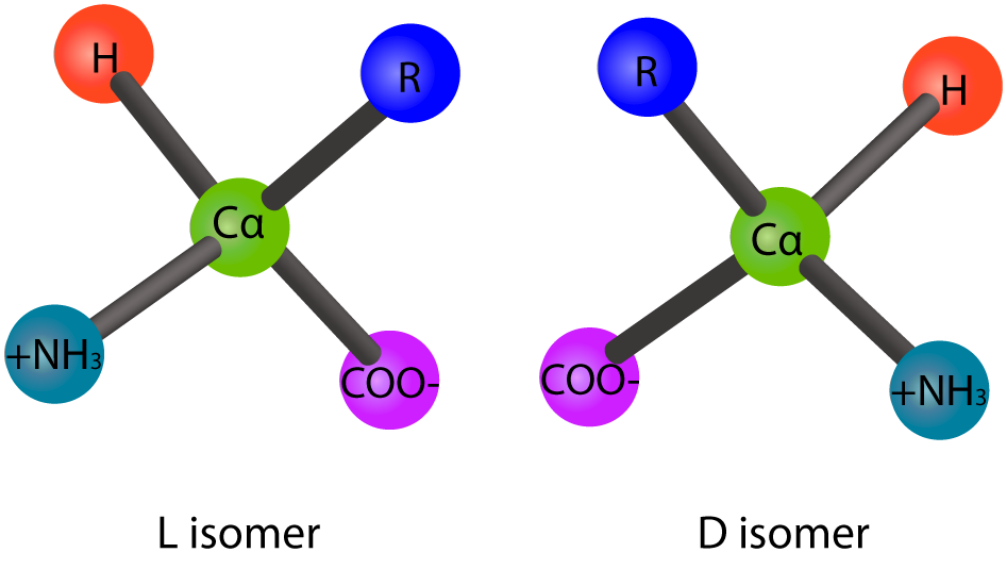


Figure 1.3 The L and D isomers of amino acids, where R refers to the unique side chain.

All amino acids found in nature are of the L isomer kind. Alcohols, thiols, thioethers, carboxylic acids and carboxamides are some of the functional groups present in the distinctive R group of amino acids. At a pH of approximately 7, the amino group is protonated and the carboxyl group is deprotonated. pKa values indicate the pH at which functional groups change protonation states. pK_a values depend on the temperature and the ionic strength of the microenvironment surrounding the functional group.

There are levels of structural complexity found in proteins. Polypeptide chains can fold into regular repeating structures known as α helices and β pleated sheets. These motifs are referred to as the secondary structure of proteins. The α helix is a coiled structure that is stabilised by intra-polypeptide hydrogen bonds, whereas β pleated sheets are stabilised by inter-polypeptide hydrogen bonding. β turns and Ω loops are responsible for linking consecutive β pleated sheets and α helices. Several secondary structures can link together to form the tertiary structure of a protein. Depending on the manner by which the amino acids have come together in the polypeptide chains of a protein, the overall tertiary structure can either be hydrophobic or hydrophilic. It is possible for proteins to possess a water-soluble surface with a nonpolar hydrophobic core. In the case of

channel proteins found in the cell membrane, the surface of the protein is made up of amino acids that contain R groups with nonpolar properties, whereas the centre of the protein that interacts with the salts as they pass through the membrane have polar character. Several tertiary structured proteins can link together to form a quaternary structure.[1]

1.3 Basic Concepts and Kinetics of Enzymes

1.3.1 Enzyme Energetics

For any chemical reaction to occur, the reactant molecules must possess a sufficient amount of energy to cross a potential energy barrier. This barrier is known as the Gibbs free energy of activation, or simply the activation energy $\Delta G^{\#}$. Consider the reaction,

$$A + B \rightleftharpoons C + D \qquad \textbf{(1.1)}$$

The change in free energy of the reaction is determined by,

$$\Delta G = \Delta G^{O} + RT\ln\frac{[C][D]}{[A][B]} \qquad \textbf{(1.2)}$$

ΔG^{0} is the standard free energy change or the energy change for the reaction under standard conditions, R is the gas constant and T is the absolute temperature. The ΔG of a reaction depends only on the free energy of the products minus the free energy of the reactants (Figure 1.4). A reaction can occur spontaneously only if the ΔG of the reaction is negative. The reaction is considered to be in a state of equilibrium if ΔG is zero. For reactions that have a positive ΔG value, an input of energy is required to drive the reaction forward.

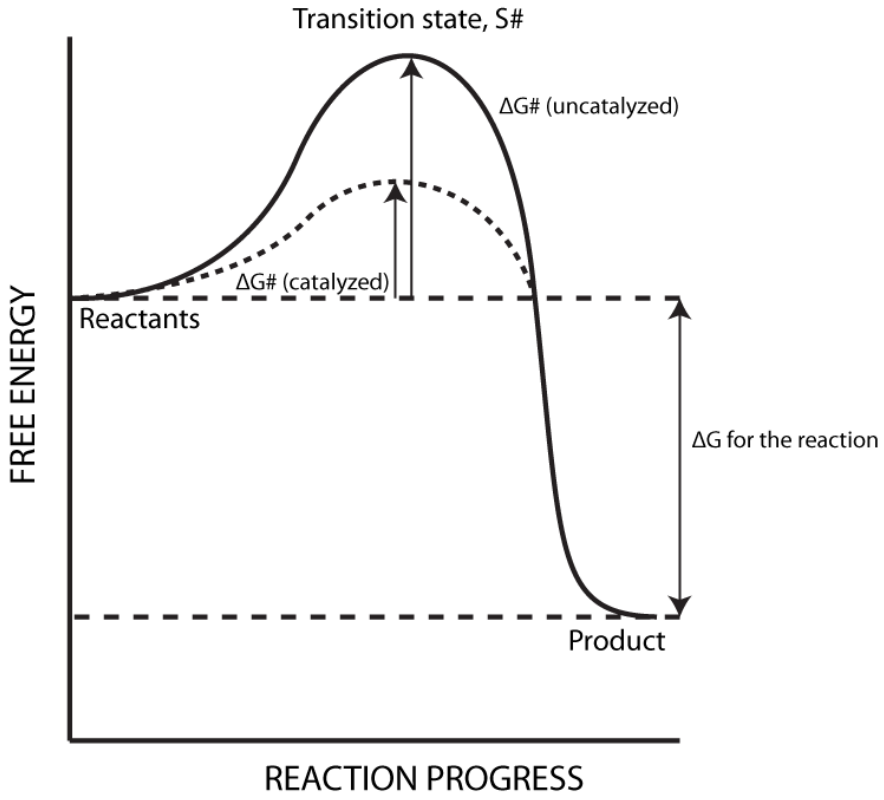


Figure 1.4 Energy profile of a reaction showing the difference between the $\Delta G^{\#}$ of an uncatalysed and catalysed reaction.

It can be seen from Figure 1.4 that ΔG is a state function which is independent of the path of the transformation from reactants to products. ΔG only indicates whether or not a reaction will occur spontaneously, it will not provide information on the rate of the reaction. The rate of the reaction depends on the activation free energy $\Delta G^{\#}$ of the reaction.

In a given chemical reaction, the reactant must go through at least one transition state in order to form the product. A transition state possesses a higher free energy than either the reactant or the product. The difference in free energy between the transition state and the reactant state is referred to as the activation energy. From the graph in Figure 1.4 it can be seen that the enzyme lowers the $\Delta G^{\#}$ by facilitating the formation of the transition state and thereby increases the rate of product formation. Consider a reactant in water compared to a reactant in the binding pocket of an enzyme. In the binding pocket of the enzyme, optimal orientation of amino acids achieves permanent dipoles which interact with the

reactant, as opposed to random polar interactions the reactant is exposed to when it is surrounded by only water molecules.

1.3.2 Enzyme Kinetics

Enzymes function by increasing the rate of a reaction. The rate of catalysis V_o, is defined as the number of moles of product formed per second. As the substrate concentration increases, V_o initially increases linearly and levels off asymptotically to a maximum velocity value V_{max} at higher substrate concentrations (Figure 1.5). V_{max}, is a result of saturation kinetics. This kinetic behaviour can be explained by the use of the Michaelis-Menten model.

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P \qquad \textbf{(1.3)}$$

In the Michaelis-Menten model, the enzyme (E) combines with the substrate (S) at a rate constant of k_1, to form an ES complex which is considered a necessary intermediate in catalysis. The ES complex can either dissociate to S and E with a rate constant of k_{-1}, or it can proceed to form the product with a rate constant of k_2. In this model however, we ignore the back reaction and assume that almost none of the product reverts to the initial substrate.

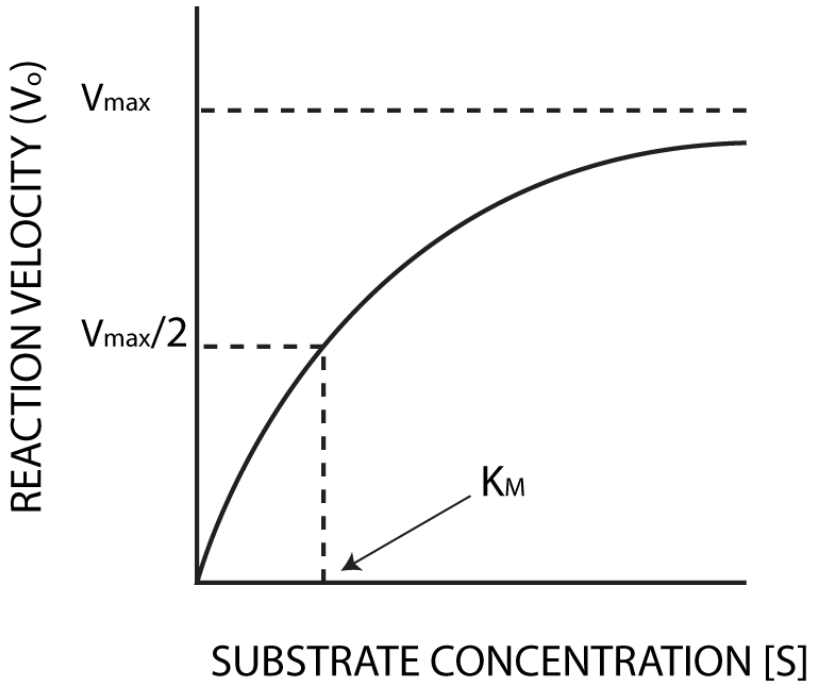


Figure 1.5 Enzyme reaction kinetics. A plot of the reaction velocity V_o as a function of substrate concentration.

The goal of the Michaelis-Menten model is to obtain an expression that relates the concentration of the substrate and the enzyme to the rate of catalysis. Due to the importance of the formation of the ES complex, the catalytic rate is dependent on,

$$V_0 = k_2[ES] \quad \textbf{(1.4)}$$

where the rate of formation of ES is,

$$ES = k_1[E][S] \quad \textbf{(1.5)}$$

and the rate of breakdown of ES is,

$$ES = (k_{-1} + k_2)[ES] \quad \textbf{(1.6)}$$

Assuming steady-state conditions for the concentration of ES, the Michaelis constant K_M, is obtained.

$$K_M = \frac{k_{-1} + k_2}{k_1} \quad \textbf{(1.7)}$$

The concentration of uncombined enzyme [E] is given by,

$$[E] = [E]_T - [ES] \quad \textbf{(1.8)}$$

where $[E]_T$ is the total enzyme concentration. Rearranging equation 1.7 and remembering that V_{max} is reached when the enzyme is saturated with substrate,

$$V_{\max} = k_2[E]_T \quad \textbf{(1.9)}$$

the Michaelis-Menton equation is attained.

$$V_0 = V_{\max}\frac{[S]}{[S] + K_M} \quad \textbf{(1.10)}$$

From equation 1.10 it can be seen that at a low substrate concentration the rate is directly proportional to the substrate concentration, whereas, at high substrate concentration $V_0=V_{max}$. If $K_M=[S]$, then $V_0=V_{max}/2$. Therefore it can be seen that K_M is the substrate concentration at which the reaction rate is half the maximum rate. The Michaelis-Menten model is used in the study of all enzyme kinetics and is very useful in understanding the nature of binding between the enzyme and the substrate.[1]

1.3.3 Enzyme Inhibition

Enzymes can encounter molecules that prevent or inhibit the enzyme from catalysing a reaction. Molecules of this nature are referred to as inhibitors and may serve as a type of control mechanism in biological systems. Inhibitors can also be used to study the binding pocket of the enzyme and allow for the identification of important catalytic residues. The extent of inhibition can either be reversible or irreversible. Irreversible inhibition can occur when the inhibitors form stable covalent or non-covalent interactions. Dissociation during irreversible inhibition can occur very slowly from the target enzyme. Reversible inhibition differs from irreversible inhibition in that dissociation occurs much faster due to weaker interactions and the absence of covalent bonding.

There are two types of reversible inhibition that can occur. In competitive inhibition, the inhibitor competes with the natural substrate for the binding pocket of the enzyme. An enzyme can bind the substrate or the inhibitor, but it cannot bind both simultaneously. Competitive inhibition effectively reduces the amount of enzyme available to bind with the substrate and therefore decreases the rate at which the product is formed. In non-competitive inhibition, the inhibitor and the natural substrate can bind to the enzyme simultaneously at different respective binding pockets. This kind of inhibition reduces the product turnover number rather than reducing the amount of enzyme available to bind with the natural substrate.

Non-competitive and competitive inhibition can be kinetically distinguished. In Figure 1.6(A), non-competitive inhibition lowers the value of V_{max} while not affecting the value of K_M. Since the enzyme-inhibitor-substrate complex does not proceed to form a product, a non-competitive inhibitor causes the remaining enzyme to behave as a more dilute solution of the functional enzyme. In Figure 1.6(B), competitive inhibition lowers the K_M value, however the V_{max} value remains unchanged. Competitive inhibition can be overcome by a sufficiently high concentration of substrate that will "outcompete" the inhibitor for the binding pocket of the enzyme.[1]

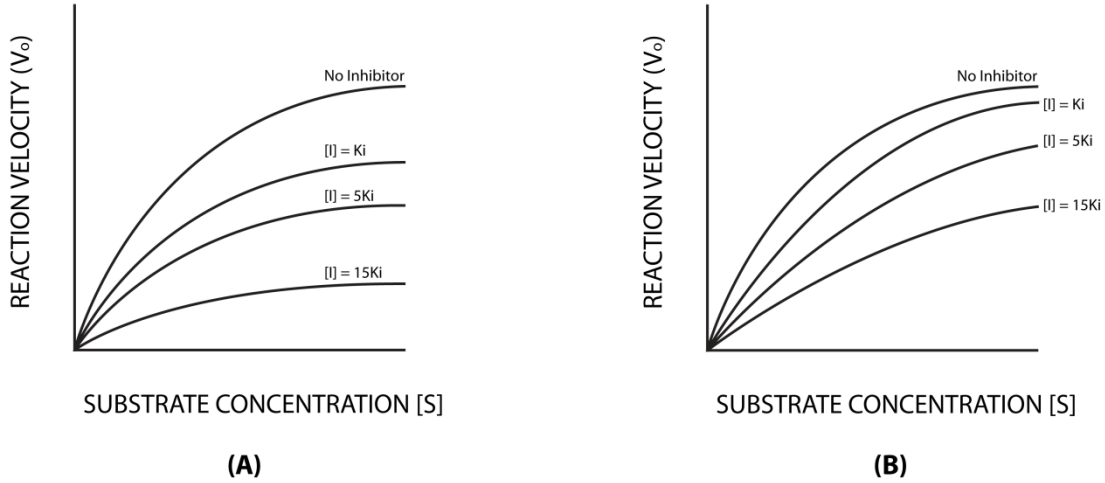


Figure 1.6 Reversible inhibition can be divided into (A) non-competitive and (B) competitive inhibition.

1.4 Enzyme Classification

Enzymes are classified according to the type of reaction they carry out. There are six classes of enzyme.

1. Oxidoreductases:
   Oxidation-reduction reactions.

2. Transfererases:
   Responsible for transferring atoms or functional groups such as amino, acetyl, phosphate and methyl groups between two molecules.

3. Hydrolases:
   Catalyse the hydrolytic cleavage of C-O, C-N, C-C and a few other bonds, including phosphoric anhydride bonds.

4. Lyases:
   Cleave bonds via an elimination reaction to form a double bond.

5. Isomerases:
   Carries out the rearrangement of atoms within a molecule.

6. Ligases:
   Joins two molecules together at the expense of ATP hydrolysis.

1.5 Base Excision Repair Enzymes

The cellular genome is constantly mutating. The high frequency with which these mutations occur is not compatible with sustaining life. Therefore corrective measures have evolved to repair these mutations. One such measure is the base excision repair pathway, shown in Figure 1.7, taken from Friedberg et al[2]. DNA N-glycosylases are the primary enzymes responsible for this pathway. They hydrolyse the N-glycosidic bond between the base and the sugar. The site where the base is removed is called the AP site because it is devoid of a purine or a pyrimidine, and therefore known as apurinic or apyrimidinic. These AP sites are cytotoxic and mutagenic and have to be further processed through the addition of the correct base. Some DNA glycosylases have associated AP lyase activity, or an AP-endonuclease cleaves the phosphodiester bond. The remaining phosphate residue is cleaved by a phosphodiesterase which liberates the sugar. The resulting vacancy site is filled with the correct sugar and base combination by a DNA polymerase enzyme and bonded to the phosphate backbone via DNA ligase.[3, 4]

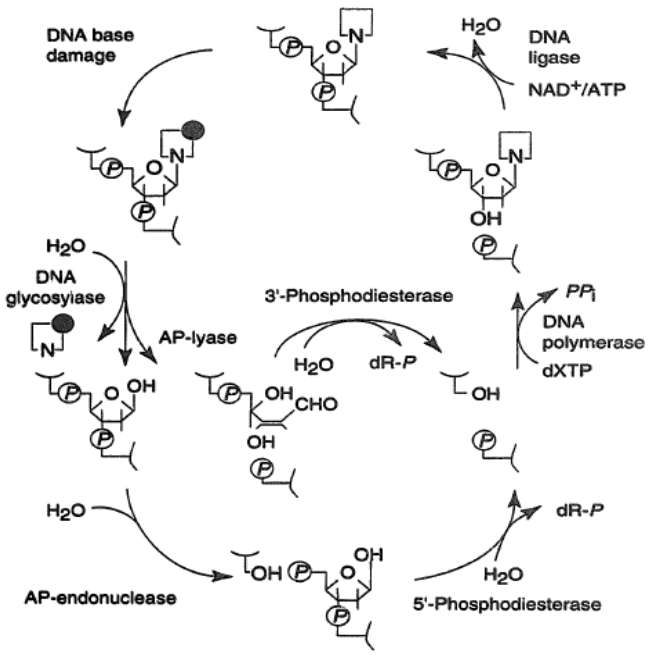


Figure 1.7 The base excision repair pathway.[4-6]

1.6 Uracil-DNA Glycosylase Superfamily

DNA glycosylases are responsible for the excision of damaged or foreign bases in DNA and is the first step in the initiation of the base excision repair pathway. The uracil-DNA glycosylase (UDG) enzyme removes uracil from DNA by the cleavage of the glycosidic linkage between the base and the sugar ring in the nucleotide (Figure 1.8).[7]

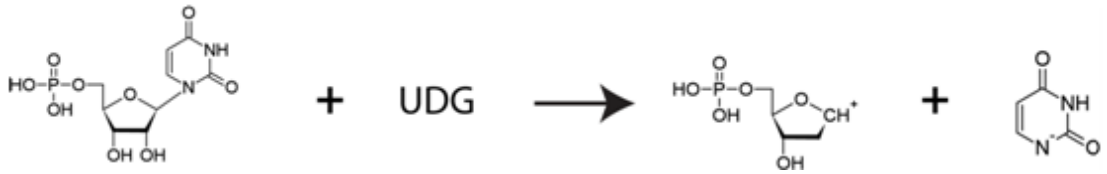


Figure 1.8 Uracil-DNA glycosylase cleaves the N-glycosidic linkage in the uracil nucleotide.

Uracil is not used in storing genetic information in DNA due to the high frequency with which uracil mutations occur. A common way in which uracil is formed in DNA is by cytosine deamination by water to produce guanine-uracil base pairs. Uracil is also present in DNA due to the misincorporation of

deoxyuridine triphosphate (dUTP) instead of deoxythymidine triphosphate (dTTP) by DNA polymerase during DNA replication. Uracil is found naturally in RNA and forms stable hydrogen bond interactions with adenine in DNA just as it does with adenine in RNA.[2]

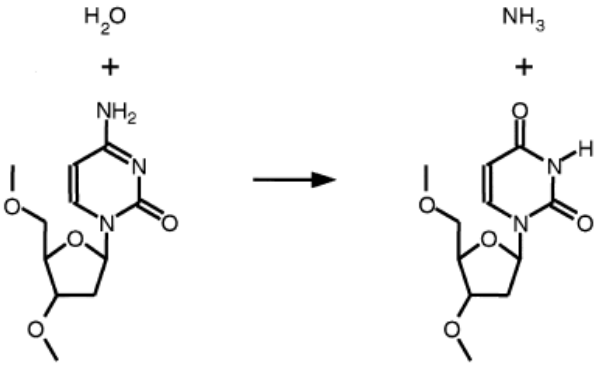


Figure 1.9 The deamination of cytosine to produce uracil.

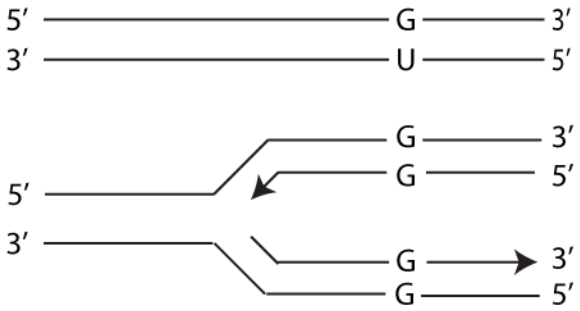
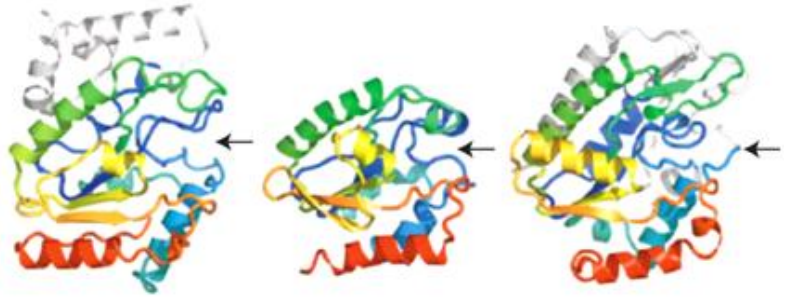


Figure 1.10 Unchecked uracil mutations lead to A:U mismatches in DNA. 5' and 3' indicate the trailing and leading carbons of the sugar ring connected along the sugar phosphate backbone respectively.

If this mutation was allowed to proceed uncorrected, adenine-uracil mutations will be produced in half of the daughter duplexes after DNA replication (Figure 1.10).

The UDG superfamily is divided into families based on conserved active site residues and the specificity of the enzymes. They are monofunctional BER enzymes that occur in viruses, prokaryotes and eukaryotes and are divided into five families. Family-1 UDG is the most well understood of the five families.[6]

Family-1 UDG is specific for uracil regardless of its base paring partner. UDG may remove uracil from either double or single stranded DNA. Uracil nucleotides are "flipped" into the active site of the enzyme which has a high affinity for the AP site of the DNA. Base flipping refers to the process by which the base flips out from the centre of the DNA double helix and assumes the extrahelical conformation. Initially it was thought that UDG enzymes scanned the entire length of the DNA strand sampling each base and forming only stable interactions when it encountered a uracil base.[8, 9] Recent studies have however determined that base flipping is a natural dynamic process.[10, 11] Thymine-uracil base pairs caused by mutation, have been found to be less stable than the naturally occurring base pairs. Due to this, they are more likely to exist in the extrahelical position. DNA helix stability is also dependent on the composition of bases found in the DNA. Due to these aforementioned findings, it has been proposed that the UDG enzyme captures the uracil base as it spontaneously flips out of the DNA double helix, where it may partition forward into the enzyme active site, or back into the DNA double helix.[12, 13]



(A) (B) (C)

Figure 1.11 Secondary structure of (A) Family-1 herpes simplex virus type 1 UDG, (B) Family-2 E. coli mismatch UDG and (C) Family-3 Xenopus single strand selective mono-functional UDG. The arrows indicate the active sites.[14, 15]

The secondary structure of the UDG enzyme across the UDG superfamily can be seen in Figure 1.11 The N-terminal is shown in blue and the O-terminal of the protein is shown in red. All the enzymes possess the conserved active site in relatively the same position.[16, 17] The active sites of the enzymes show signs of

inherited amino acids across the superfamily. The green amino acid in Figure 1.12 represents phenylalanine which forms the side wall of the pocket and the blue amino acids represent glycine and proline which form part of the catalytic motif. The underlined amino acids in Figure 1.12 represent the conserved catalytic residues.

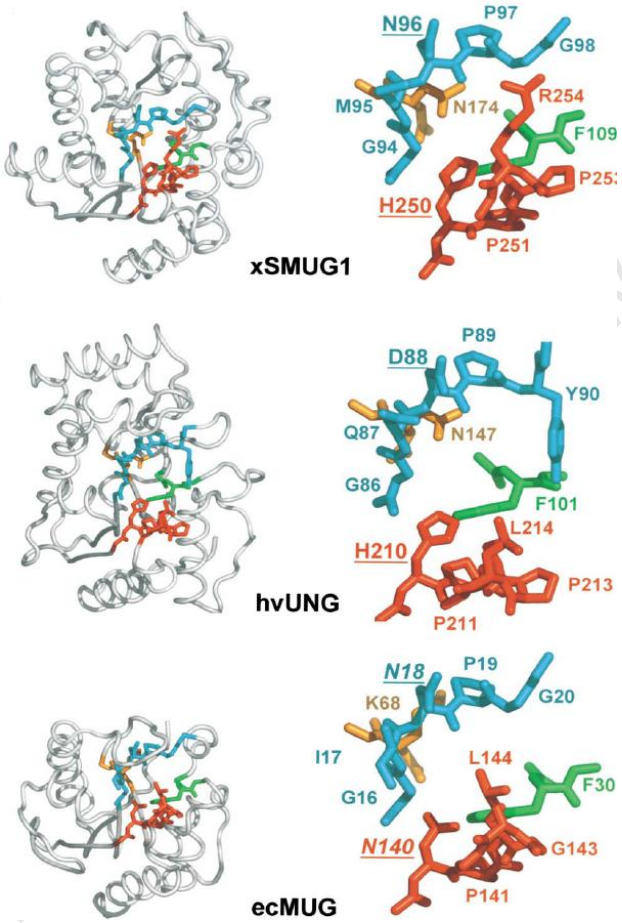
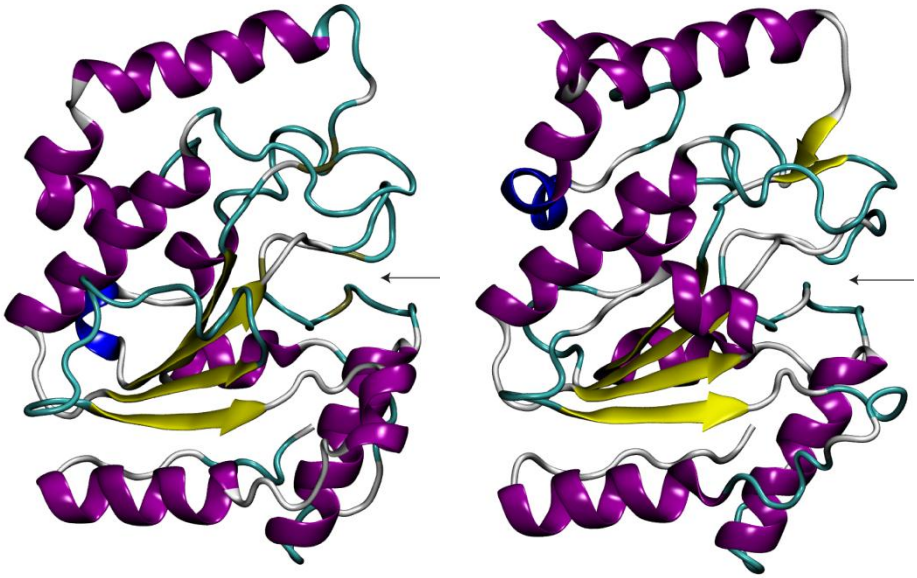


Figure 1.12 Active site of UDG across the superfamily. Conserved residues are underlined.[15]

1.7 The Conserved Nature of Human Uracil-DNA Glycosylase and Herpes Simplex Virus Type 1 Uracil-DNA Glycosylase enzymes.

Crystal structures of human uracil-DNA glycosylase (hUDG)[8] and herpes simplex virus type 1 uracil-DNA glycosylase (hsvUDG)[16] have been produced with resolutions of 1.90Å and 1.75Å. The overall secondary structure of both enzymes consists primarily of 8 α helices and 1 β pleated sheet. The active site is situated

between a loop, indicated by the arrow in Figure 1.13. Due to evolutionary similarities, there are a number of conserved catalytic amino acids present in the binding pockets of these proteins. The buried uracil-binding pocket is characterised by a general shape and complementary electrostatics to the 2-, 3- and 4-positions of uracil.



(A) **(B)**

Figure 1.13 Secondary structure of (A) human and (B) herpes simplex virus type 1 Uracil-DNA glycosylase. Helices are displayed in purple, loops in green and sheets in yellow. The arrows indicate the active site of the enzymes.

The primary goal of uracil-DNA glycosylase enzymes is to facilitate the cleavage of the glycosidic bond connecting the deoxyribose sugar ring to the uracil base.[18] In enzymatic glycosyl transfers such as this one, unstable glycosyl cation transition states are formed. These transition states require stabilisation by the active site of the enzyme. The proposed reaction (Figure 1.14) produces an oxocarbenium ion intermediate. The sugar ring becomes positively charged, whereas the uracil ring becomes negatively charged. In order for the reaction to proceed, these intermediates have to be stabilised. This is facilitated by aspartic acid located at position 145 in hUDG (ASP145) and 88 in hsvUDG (ASP88) and histidine located at position 268 in hUDG (HIS268) and position 210 in hsvUDG (HIS210). These amino acid residues interact with the oxocarbenium ion

intermediate through electrostatic interactions. Histidine residues (HIS268 and HIS210) serve to stabilise the negative charge on the uracil ring by donating a hydrogen bond to the O2 oxygen of uracil. The cationic sugar ring is stabilised by aspartic acid (ASP145 and ASP88), which activates a water molecule by positioning it in such a way that the oxygen of the water molecule stabilises the positive charge on the C1 carbon of the sugar ring.[19]

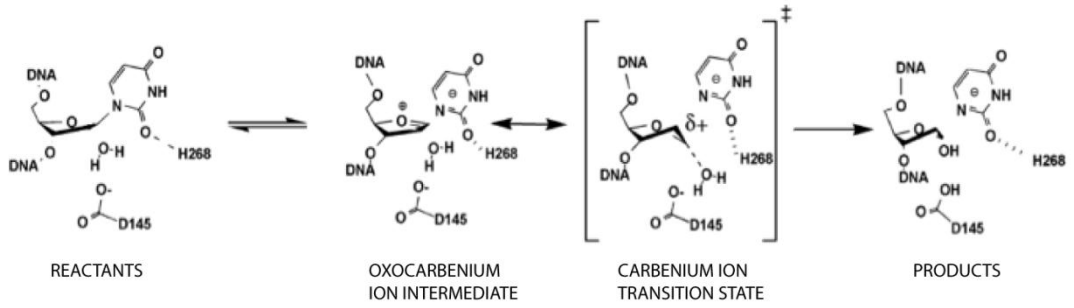


Figure 1.14 Proposed mechanism for the glycosidase activity in hUDG. Due to the highly conserved active site in family-1 UDG, this reaction can be proposed for hsvUDG.[19]

This reaction is a good example of product-assisted catalysis, which, in this situation entails the complementary stabilisation of the cationic sugar ring by the anionic uracil ring. It was also determined that UDG functions by the use of a mechanism known as "substrate autocatalysis". In this mechanism, the burial and positioning of the 4 phosphate groups of the DNA base pair being corrected assists in stabilising the transition state. According to computational studies carried out on this enzyme, the amino acids, HIS268 and ASP145 contribute 6.9 kcal.mol^{-1} to lowering the activation energy, whereas the 4 phosphate groups contribute 21.9 kcal.mol^{-1}, which is considerably larger than that of the amino acid residues. The protein backbone is also said to provide an additional 4.7 kcal.mol^{-1}.

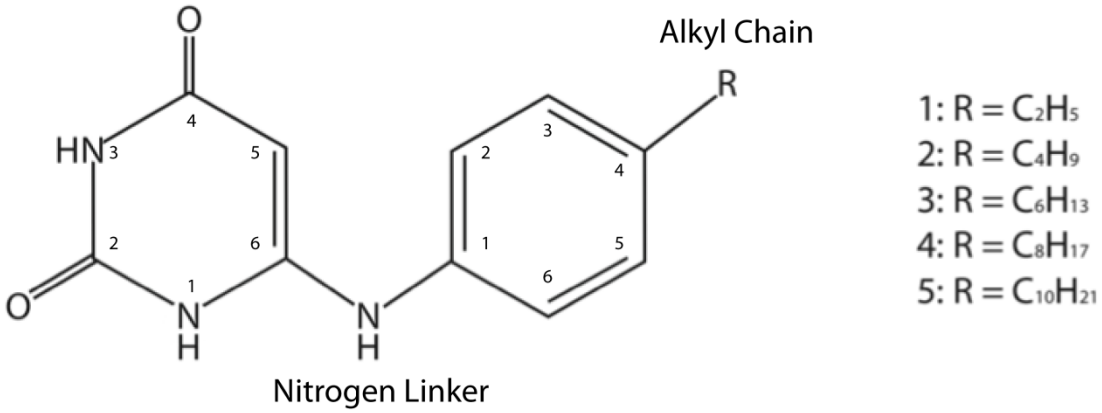
Mutation studies have been carried out on the hUDG enzyme. Due to the highly conserved nature of the proteins, the findings of these studies can be extended to the hsvUDG enzyme. These studies revealed that a leucine residue located at position 272 in hUDG (LEU272) and 214 in hsvUDG (LEU214) is important as it

penetrates the DNA base stack, thereby replacing the flipped-out uracil nucleotide. Mutating LEU272 into an alanine residue leads to severely impaired uracil excision.[8] Catalytic amino acids in the binding pocket are responsible for the specific recognition of uracil. Mutating the asparagine located at position 204 in hUDG (ASN204) and 147 in hUDG (ASN147) to an aspartic acid amino acid leads to the enzyme having cytosine-DNA glycosylase activity. Tyrosine located at position 147 (TYR147) has been determined to be important for the discrimination of uracil over thymine. When TYR147 is mutated into alanine in hUDG, the enzyme receives thymine-DNA glycosylase activity.[20] These catalytic amino acids are very important in selecting uracil over the other bases and as a result, are conserved throughout the uracil-DNA glycosylase family.

Theoretical[21] and experimental[22, 23] studies indicate that the phosphate groups which form part of the DNA backbone, provide a significant amount of electrostatic interactions. This allows for further stabilisation of the DNA substrate in the binding pocket.[24] These interactions can occur through direct amino acid hydrogen bond formation or through contacts mediated by water.[25, 26]

1.8 Selective inhibition of Herpes Simplex Virus Type 1 Uracil-DNA Glycosylase.

Studies[27] have shown that the herpes simplex virus type 1 (HSV1) requires uracil-DNA glycosylase (hsvUDG) for reactivation following a latency period in the life cycle of the virus. During the latency stage of the virus, its DNA undergoes various mutations. hsvUDG is responsible for postreplicative DNA repair by removal of uracil residues from the DNA. 6-(4-alkylanilino)-uracil molecules were found to be effective inhibitors of the hsvUDG enzyme.[28] Effective inhibition of the hsvUDG enzyme is achieved at a micro molar (μM) range without effectively inhibiting the hUDG enzyme (Figure 4). All inhibitors for the hUDG enzyme produced IC_{50} values of greater than 500μM.



(A)

Inhibitor	IC_{50} (μM)
1	500
2	150
3	30
4	8
5	35

(B)

Figure 1.15 The (A) molecular structures of the 6-(4-Alkylanilino)-uracil inhibitor sand their respective (B) IC_{50} values.[28]

The 6-(4-octylanilino)uracil, inhibitor 4, was found to possess the best binding affinity for the hsvUDG enzyme. According to modelling and structural design studies, inhibitor 4 possesses 5 freely rotatable single bonds making it highly flexible. The 6-NH bond is approximately perpendicular to the phenyl ring of TYR 90 at a distance of 3.85 Å which allows for an interaction energetically equivalent to about one-half of a normal hydrogen bond to form. PRO111, PRO213 and LEU214 provide a hydrophobic cleft that the second half of the octyl chain can firmly be placed in.[28] These non-covalent interactions between the enzyme and ligand are crucial in affinity models.[28]

1.9 Objectives

The primary objective of this work is to rationalise the inhibitory effects of the 6-(4-Alkylanilino)-uracil inhibitors and to gain insight into the selective nature of these inhibitors for the herpes simplex virus uracil-DNA glycosylase enzyme. This can only be achieved by observing the interaction between the protein and the substrates in a solvated environment. From these observations, structural changes as well as important amino acids, and correlations in energy changes and interaction distance, can be identified that will allow for further development of inhibitors.

Free energy perturbation (FEP) and molecular dynamics will be used in this thesis. Force field parameters for inhibitors will be created and then optimised. FEP simulations and experimental data will then be used to validate these parameters. Simulation will be performed using DNA as a substrate for both the hsvUDG and hUDG enzymes using periodic boundary conditions. Simulations for the hsvUDG and hUDG enzymes with all inhibitors will be performed using stochastic boundary condition. These simulations will provide insight into the selective nature of the inhibitors.

References:

1. J. M. Berg, J. L. Tymoczko and L. Stryer, *Biochemistry*, 5th edn., Freeman and Company, New York, 2002.
2. E. C. Friedberg, G. C. Walker and W. Siede, *DNA repair and mutagenesis*, ASM Press Washington, DC., 1995.
3. H. E. Krokan, R. Standal and G. Slupphaug, *Biochemical Journal*, 1997, **325**, 1-16.
4. D. O. Zharkov, *Cellular and Molecular Life Sciences*, 2008, **65**, 1544-1565.
5. E. Seeberg, L. Eide and M. Bjoras, *Trends in Biochemical Sciences*, 1995, **20**, 391-397.
6. P. J. Berti and J. A. B. McCann, *Chemical Reviews*, 2006, **106**, 506-555.
7. T. Lindahl, *Proceedings of the National. Academy of Sciences*, 1974, **71**, 3649-3653.
8. S. S. Parikh, C. D. Mol, G. Slupphaug, S. Bharati, H. E. Krokan and J. A. Tainer, *Embo Journal*, 1998, **17**, 5214-5226.
9. G. Slupphaug, C. D. Mol, B. Kavli, A. S. Arvai, H. E. Krokan and J. A. Tainer, *Nature*, 1996, **384**, 87-92.
10. C. Cao, Y. L. Jiang, J. T. Stivers and F. Song, *Nature Structuraland Molecular Biology*, 2004, **11**, 1230-1236.
11. I. Wong, A. J. Lundquist, A. S. Bernards and D. W. Mosbaugh, *Journal of Biological Chemistry.*, 2002, **277**, 19424-19432.
12. S. R. W. Bellamy, K. Krusong and G. S. Baldwin, *Nucleic Acids Research*, 2007, **35**, 1478-1487.
13. C. Cao, Y. L. Jiang and J. T. Stivers, *Journal ofthe American Chemical Society*, 2006, **128**, 13034-13035.
14. J. E. A. Wibley, T. R. Waters, K. Haushalter, G. L. Verdine and L. H. Pearl, *Molecular Cell*, 2003, **11**, 1647-1659.
15. L. H. Pearl, *Mutation Research-DNA Repair*, 2000, **460**, 165-181.
16. R. Savva, K. McAuleyhecht, T. Brown and L. Pearl, *Nature*, 1995, **373**, 487-493.
17. C. D. Mol, A. S. Arvai, G. Slupphaug, B. Kavli, I. Alseth, H. E. Krokan and J. A. Tainer, *Cell*, 1995, **80**, 869-878.
18. A. R. Dinner, G. M. Blackburn and M. Karplus, *Nature*, 2001, **413**, 752-755.
19. R. M. Werner and J. T. Stivers, *Biochemistry*, 2000, **39**, 14054-14064.
20. B. Kavli, G. Slupphaug, C. D. Mol, A. S. Arvai, S. B. Peterson, J. A. Tainer and H. E. Krokan, *Embo Journal*, 1996, **15**, 3442-3447.
21. A. Ma, J. Hu, M. Karplus and A. R. Dinner, *Biochemistry*, 2006, **45**, 13687-13696.
22. Y. L. Jiang and J. T. Stivers, *Biochemistry*, 2001, **40**, 7710-7719.
23. Y. L. Jiang, Y. Ichikawa, F. Song and J. T. Stivers, *Biochemistry*, 2003, **42**, 1922.
24. M. Olufsen, A. O. Smalas and B. O. Brandsdal, *Journal of Molecular Modeling*, 2008, **14**, 201-213.
25. Z. Moravek, S. Neidle and B. Schneider, *Nucleic Acids Research*, 2002, **30**, 1182-1191.
26. G. A. Nevinsky, *Molecular Biology*, 2003, **38**, 636-662.

27. A. Verri, P. Mazzarello, G. Biamonti, S. Spadari and F. Focher, *Nucleic Acids Research.*, 1990, **18**, 5775-5780.
28. H. M. Sun, C. X. Zhi, G. E. Wright, D. Ubiali, M. Pregnolato, A. Verri, F. Focher and S. Spadari, *Journal of Medicinal Chemistry*, 1999, **42**, 2344-2350.

Chapter 2

Essential Theory of Computational Biochemistry

2.1 Introduction

In 1977 the first simulation of a protein known as bovine pancreatic trypsin inhibitor using molecular dynamics was reported.[1] This brought about a great interest in a new field of research that would grow at a rapid rate. Crystal structures of proteins provided a static picture of biomolecules, which led to the incorrect conclusion that proteins are rigid structures and the development of theories such as the "lock and key" model of enzyme catalysis. Today we widely accept the dynamic nature of biological molecules based on experimental evidence obtained of a given system at different stages of its cycle.[2] Triosephosphate isomerase is an enzyme that has a loop consisting of 11 amino acid residues that undergoes a displacement of about 7Å after binding with its substrate.[3] This is just one example of the dynamic behaviour of biomolecules. Understanding the dynamic nature of a protein provides insight into its functionality.

Computational approaches have been used to investigate the energetics associated with conformations and chemical structure as well as to compare the relative free energy differences and barriers of molecular systems and compare them to experimental results. Using a combination of computational and experimental research methods allows one to streamline ones research method. Drug design and development is a field of research that has benefited hugely by scientific computing methods.[4, 5] Pharmacological properties can easily be adjusted and tested and thereby the refinement of drug candidates can be achieved using computational approaches.

There are many reasons for the rapid progress of computational and theoretical studies of biological molecules.[6] The growth of the protein data bank, which is a

repository of experimentally determined 3-dimensional coordinates of protein structures and molecules as well as a significant improvement in computational methodologies and increase in computational resources has established scientific computing as a compelling field of research and one that's applicability continues to grow.[7] Quantum mechanics (QM) and Molecular Mechanics (MM) are two fundamental computational approaches to solving biochemical problems. Using Molecular Dynamics (MD) methods along with MM allows for dynamic data to be evaluated. Statistical mechanics can be applied to the sampling data obtained from MD simulations to evaluate thermodynamic data of the system being simulated.

Deciding on what computational method to use depends on the requirements of the task being taken on. Factors such as time constraints, computational power, accuracy required and other limiting factors determine which method would be more practical. Table 2.1 summarises computational methods available.

	Classical Mechanics	**Quantum Mehcanics**		
	MM and MD	**ab initio**	**Semi Empirical**	**Density Functional Theory**
Description	Uses ball and spring model and newtonian equations of motion.	Solves the Schrödinger equation using 1st principles.	Solves the Schrödinger equation using theory and experimental data.	Solves Schrödinger equation and functional of electron density.
Application	Large molecular systems. Proteins and DNA.	Novel molecules and active sites of enzymes. Or reduced model systems.	Moderately large systems.	Novel Molecules.
Limitations	Force Field specific.	Computaionally demanding.	Parameterisation specific.	Computationally demanding.

Table 2.1 Summary of classical and quantum mechanical simulation methods.[8, 9]

2.2 Molecular Mechanics

For large biomolecular systems, *ab initio* or semiempirical methods are rarely applied efficiently due to the large size of such systems and the computational resources required. For optimisation of large systems, it is more practical to use

Molecular Mechanics (MM). MM uses simple algebraic expressions in order to calculate the total energy of a compound or system without having to solve a wavefunction or the electron density.[10] These algebraic expressions (potential energy functions) and the parameter constants used by these functions for evaluating interactions are collectively termed the force field.

The parameters used in the force field are derived empirically. Ignoring the electron and considering the position of the nucleus as the centre of mass of the atom simplifies the model of the molecular calculation. Using the single nuclear coordinate to represent the atom is justified in terms of the Born-Oppenheimer approximation.[11] These atom-like particles are treated as spherical balls and the bonds between them are treated as springs. Using these simple models along with Newton's equations of motion, structural fluctuations in the conformations of the atom-like particles can be observed with respect to time.[12]

2.2.1 Force Fields and Atomic Modelling

The force field or potential energy function establishes an essential link between chemical structure and energy in atomistic models of biochemical systems. Atomic coordinates are used to determine bond lengths, angles and distances between atoms. These variable are then assessed in the force field to calculate the potential energy of the system.[10] The total potential energy of a given system is expressed in the form,

$$E_{TOTAL} = E_{bonded} + E_{non} \quad (2.1)$$

where the bonded terms are,

$$E_{bonded} = E_b + E_\theta + E_\phi \quad (2.2)$$

in which the successive terms expressing the bonded energy (E_{bond}) are energies associated with bond stretching (E_b), bond angle bending (E_θ) and bond torsion (E_ϕ). The nonbonded terms are,

$$E_{nonb} = E_{vdW} + E_{el} \quad (2.3)$$

in which the van der Waals (E_{vdW}) and electrostatic interactions (E_{el}) constitute the nonbonded interactions (E_{nonb}). Bonded interactions occur between covalently connected atoms and non-bonded interactions occur between noncovalent intermolecular atoms (Figure 2.1).[8] For the CHARMM program, force fields have been separately developed for proteins, nucleic acids[13], lipids[14] and carbohydrates[12, 15].

A harmonic potential such as Hook's law is applied to the bonded interactions and the nonbonded interactions are described using simple pairwise additive functions.[16]

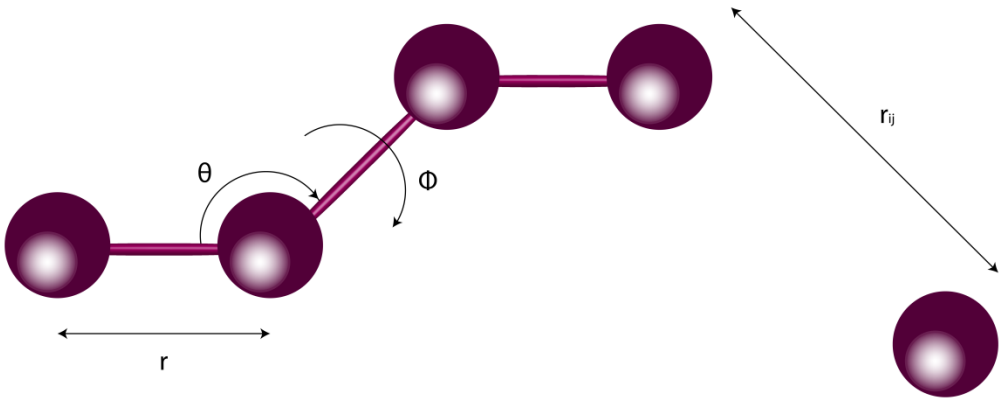


Figure 2.1 Theoretical molecules to illustrate the energy terms in equations 3.1-3.3. The values, θ, Φ, r and r_{ij} represent the angle, dihedral, bond length and interatomic distance.

If we assume that Hooke's law is adequate to describe the bond stretching between atoms, then the formula

$$E_b = \sum K_b (r - r_0)^2 \quad (2.4)$$

can be used to describe the potential energy of the bond between two atoms. The parameters K_b and r_0 are the force constant and the equilibrium bond length of the bond respectively. The instantaneous bond length is represented by the r parameter. Bond angle bending can also be described harmonically.

$$E_\theta = \sum K_\theta (\theta - \theta_0)^2 \quad (2.5)$$

In the above formula K_θ and θ_0 are the force constant and equilibrium value of the angle respectively. If we use the theoretical model in Figure 2.1, we can see that as Φ changes, the relative positions of atoms 1 and 4 can change from a low energy staggered conformation to a high energy eclipsed conformation and then back to a low energy staggered conformation. This oscillatory nature of the dihedral potential energy can accurately be modelled by the use of a sinusoidal function.

$$E_\phi = \sum K_\phi [1 - \cos(n\phi - \gamma)] \quad (2.6)$$

The parameters K_φ, γ and n represent the force constant, the phase shift and the periodicity of the dihedral being modeled. The periodicity determines the number of cycles in the oscillation per 360° rotation about the dihedral. To describe an sp^3-sp^3 such as ethane, the periodicity would be 3. However to describe an sp^2-sp^2 such as ethene, the periodicity would be equal to 2. The force constant determines the barrier to rotation of the dihedral. A dihedral that involves a double bond would generally have a greater force constant than a dihedral involving only a single bond. The phase shift of the above formula dictates the location of the maxima and minima in the dihedral energy surface.

In empirical energy functions, nonbonded interactions tend to be described using simple pairwise additive functions. In a classical world as two particles approach one another there is no attractive force between them. The potential

energy remains zero as the two atoms come together. When the distance reaches the combined length of the van der Waals radii of the two atoms ($r < 2r_{vdw}$), the potential energy discontinuously becomes infinity. This hard sphere potential energy is shown in Figure 2.2 by the solid straight lines.

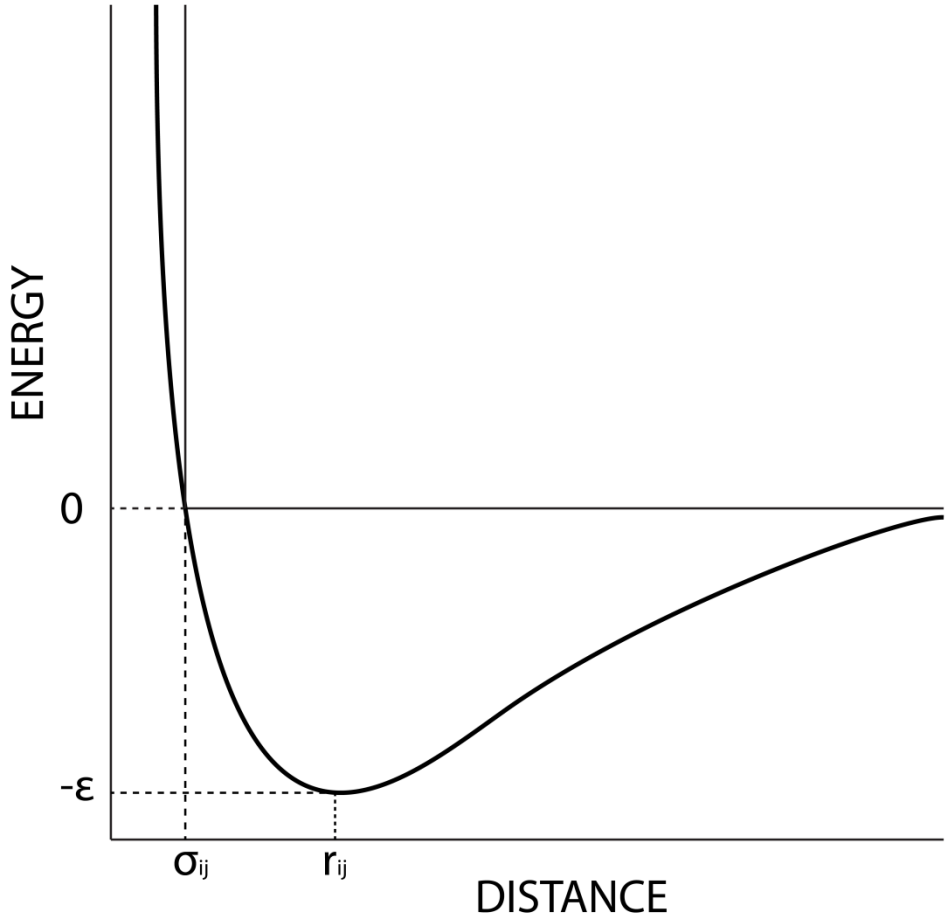


Figure 2.2 Graph of a Non-attractive hard sphere potential (straight line) and the Lennard-Jones potential (curved line).

The hard sphere model does not adequately describe what happens in reality as it neglects any form of nonbonded interaction between the atoms. The Lennard-Jones potential (equation 2.7) describes the dispersion and repulsion interactions between two atoms and represents a more accurate model for the van der Waals interactions.[8]

$$E_{vdW} = 4\varepsilon\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) \qquad (2.7)$$

The interatomic separation at which the repulsive and attractive forces cancel each other out is represented by σ_{ij}. r_{ij} represents the varying lengths between the atoms. An inversely proportional relationship exists between the electrostatic energy and the distance between the atoms. Assigning a partial charge to each atom *i* and *j*, the energy can be described by using Coulomb's law,

$$E_{el} = \sum_{i>j} \frac{q_i q_j}{4\pi\varepsilon_o r_{ij}} \qquad (2.8)$$

where r_{ij} is the separation distance between atoms *i* and *j* and q_i and q_j are the partial charges of the atoms. $4\pi\varepsilon_0$ in the equation is the permittivity constant which relates electric charge to mechanical quantities such as length and force.

2.2.2 Energy Minimisation

In nature, biological molecules prefer to be in a state of stability. Molecules will arrange in such a way to reduce inter or intra molecular steric hindrance and thereby minimise the energy of the molecule. This rearrangement of the molecule can be mathematically interpretedin molecular mechanics as the minimisation of the potential energy function. Minimisation involves the calculation of the first and second derivative of the potential energy function with respect to the coordinates (*r*), and determining at which of the variables is the first derivative equal to zero and second derivative positive[17] (equation 2.9).

$$\frac{\partial f}{\partial r} = 0; \qquad \frac{\partial^2 f}{\partial^2 r} > 0 \qquad (2.9)$$

Due to the manner by which the energy varies because of the change in coordinates of a molecular system, it is not always possible to simply minimise a function of a molecular system in one step (Figure 2.3). Instead, there are iterative algorithms designed to locate the minimum energy of a molecular

system by gradually changing the coordinates of the system. With each iterative step, the energy of the system is lowered until the minimum energy is reached. Several commonly used minimisation algorithms use derivative techniques. The steepest descent and conjugate gradient algorithms use a first order derivative scheme whereas the Newton-Raphson algorithm uses a second order derivative scheme.[18]

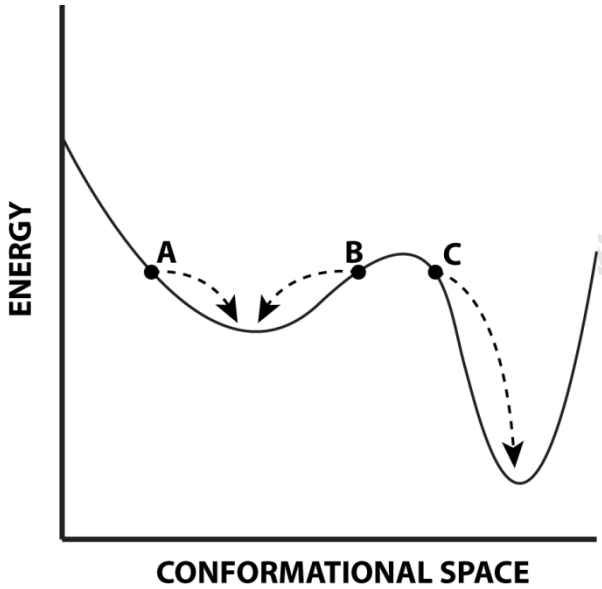


Figure 2.3 A one dimensional energy surface. A graph illustrating the change in energy of a molecule as its conformational space changes.

Energy minimisation methods are very useful and are used very often when analysing the potential energy surface of a molecule. Starting coordinates for biomolecular systems obtained from x-ray crystal structures are usually are often minimised after they have been protonated and solvated before they are simulated using molecular dynamics.

2.3 Molecular Dynamic Methods

The primary goal of molecular mechanics is the prediction of a local minimum on a molecular potential energy surface. In reality, molecules are dynamic. The continuous vibrations and rotations experienced by proteins and nucleic acids are important for biological functions. Molecules are quantum mechanical entities that are best described by the time-dependent Schrödinger equation which is the quantum mechanical equation of motion. Due to the extreme difficulty in solving this equation for large systems, a simpler description is used called molecular dynamics. Molecular dynamics is a technique used for the study of the natural time evolution of a system. This allows for the prediction of structural dynamics properties of the system by numerically solving the equations of motion. These equations of motion govern the manner by which molecules or particles in a system move and interact with each other. Because of the manner by which molecules move and interact, thermodynamic data can be extrapolated using statistical mechanics. The forces acting on the atoms in a system, which are required to simulate their movement, are calculated using the force field chosen for the simulation. The first step in the simulation of a system using molecular dynamics requires the force acting on each atom to be determined.

2.3.1 Newton's Equations of Motion

Consider a particle *A*, that has a position vector r_A and whose mass is m_A. The force F_A acting on this particle would be described by Newton's 2nd law,

$$F_A = m_A \frac{\partial^2 r_A}{\partial t^2} = m_A a_A \tag{2.10}$$

where a_A is the acceleration of particle *A*. In order to determine the force acting on particle *A* by the rest of the system, the potential energy *U*, of the system acting on the particle with respect to the position of the particle has to be determined. This potential energy is calculated by using the potential energy

function discussed in section 2.1. The negative gradient of the potential-energy function with respect to the position of particle *A*,

$$F_A = -\frac{\partial U}{\partial r_A} \tag{2.11}$$

provides the force F_A of the system acting on particle *A*. Once the force is determined, using equation 2.11 we can determine the acceleration. The integral of the acceleration over time amounts to the change in velocity *v*. The relationship between the velocity and the momentum p_A, of particle *A* is

$$v_A = \frac{\partial r_A}{\partial t} \tag{2.12}$$

$$\frac{\partial r_A}{\partial t} = \frac{p_A}{m_A} \tag{2.13}$$

By utilising the equations of motion, it is possible to calculate the position, velocity, acceleration, momentum and force acting on particle *A*. By applying the above mentioned method to atoms in a biomolecular system, the dynamic nature of the system can be evaluated.[19, 20]

2.3.2 Algorithms for Time Dependence

Solving Newton's equation of motion requires a numerical procedure for iteratively integrating the classical equations of motion for every explicit atom in a system by moving forward in time by tiny time increments, *Δt*. There are several algorithms in existence designed to carry out this iterative procedure.[21] Many of these algorithms are derived from approximate Taylor expansions such as the following expansion in *r*, the position vector,

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{1}{2!}a(t)\,\Delta t^2 + \frac{1}{3!}\frac{d^3 r(\tau)}{d^3 t}\bigg|_{\tau=t} (\Delta t)^3 + . \tag{2.14}$$

where *v(t)* and *a(t)* represent the velocity and acceleration as a function of time. Other sophisticated integration schemes that have been developed include the Leapfrog[22] and Verlet[23] integrators. The more common of the two integration algorithms used in the study of biomolecules is the Verlet integrator, which is based on two Taylor expansions of *r*, in the forward expansion (*t+Δt*)

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 + \ldots \quad (2.15)$$

and the backward expansion (*t-Δt*),

$$r(t-\Delta t) = r(t) - v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 + \ldots \quad (2.16)$$

which when summed together produces,

$$r(t+\Delta t) + r(t-\Delta t) \approx 2r(t) + a(t)\Delta t^2$$

This summation assumes that third-order and higher terms in the Taylor expansion are negligible. The acceleration is obtained from the force using equation 2.10. The Verlet algorithm uses position and acceleration at time *t* and position at time (*t-Δt*) to calculate the new position at time (*t+Δt*) for each iteration of the time-step.[9]

The Leapfrog integrator is based on two Taylor expansions of *v*, in the forward expansion (*t+Δt/2*),

$$v\left(t+\frac{\Delta t}{2}\right) = v(t) + a(t)\left(\frac{\Delta t}{2}\right) + \frac{1}{2}\left(\frac{\partial a}{\partial t}\right)\left(\frac{\Delta t}{2}\right)^2 \quad (2.17)$$

and backward expansion (t-Δt/2),

$$v\left(t-\frac{\Delta t}{2}\right) = v(t) - a(t)\left(\frac{\Delta t}{2}\right) + \frac{1}{2}\left(\frac{\partial a}{\partial t}\right)\left(\frac{\Delta t}{2}\right)^2 \quad (2.18)$$

After subtracting and rearranging,

$$v(t+\frac{\Delta t}{2}) = v(t-\frac{\Delta t}{2}) + a(t)\Delta t \tag{2.19}$$

Again the acceleration is calculated from the force. The same procedure is done for the Taylor expansion of *r* at time point *t+Δt*,

$$r(t+\Delta t) = r(t) + v(t+\frac{\Delta t}{2})\Delta t \tag{2.20}$$

Equations 2.19 and 2.20 combined form the Leapfrog algorithm that is considered the more accurate and stable technique for use in molecular dynamics. This is referred to as the "Leapfrog" because the velocities are first calculated at time *t+Δt/2*, this velocity is then used to calculate the position of the particles at time *t+Δt* for every iteration of the time-step. In other words, the velocities leap over the positions and then the positions leap over the velocities.

2.4 Simulations Environment

2.4.1 Boundary Details and Potential Energy Truncation Techniques

As a result of the computational limitations placed on simulations of biological systems, the size of the system being investigated has to be small enough so as to decrease the amount of computational time required for the completion of the simulation. By reducing the size of the system, surface effects may dominate the properties of the system. This is a crucial obstacle in realistically representing the electrostatic interactions due to the r^{-1} dependence in Coulomb's law. Several methods can be used to avoid these cluster artefacts. Periodic boundary conditions and stochastic boundary condition are two such methods.

2.4.1.1 Periodic Boundary Conditions

Figure 2.4 (A) shows a box of atoms that needs to be simulated. Two immediate problems can be identified from this situation.

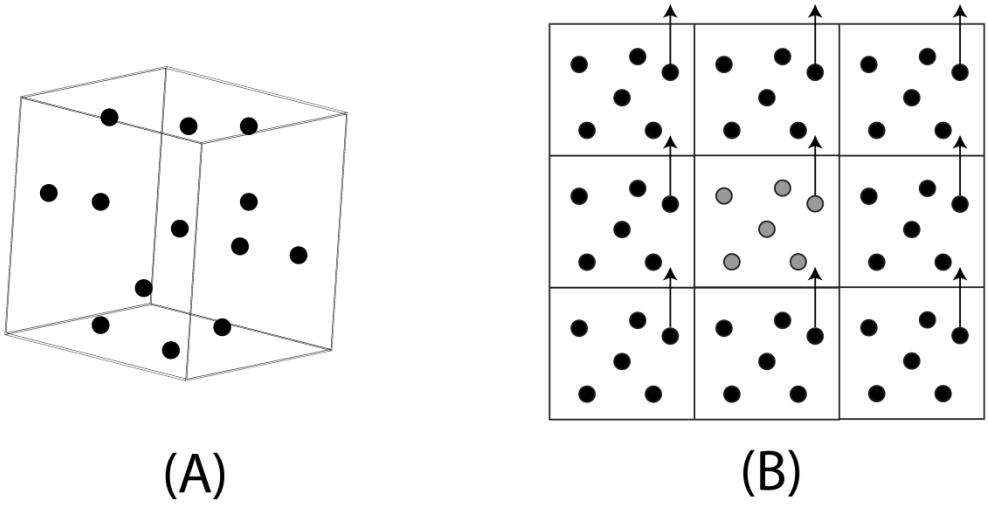


Figure 2.4 (A) Box of atoms (B) Box with periodic boundary conditions

Atoms near the edge of the box will experience different resultant forces (surface effects) than the atoms located near the centre of the box. This is because the atoms in the centre are surrounded by more neighbouring atoms. Another problem arises when the simulation forces an atom outside the boundary of the box. If this was to occur in the situation presented in Figure 2.4 (A), the density of the system would be altered.

To prevent these problems from happening, the system can be surrounded by a large number of identical copies of itself. In Figure 2.4 (B), which is a 2-dimensional slice through a small portion of this new system, the central box identified by the grey atoms, is surrounded by identical systems at the atomic level. If a dark grey atom was to leave the central cell, its image would enter from the adjoining identical system, thereby maintaining the density of the central system. This eliminates the boundary conditions presented by the wall of the box system, because each atom in the central cell is under the influence of every

other atom in the central cell and by the atoms in adjoining cells. When deciding on the box dimensions, it is important to account for the minimum image convention. This ensures that the box is large enough so that each atom does not see its image in the adjoining box.[8, 19]

2.4.1.2 Stochastic Boundary Conditions

Simulations using periodic boundary conditions usually incorporate much more solvent molecules than are actually required. When simulating and investigating the binding pocket of a substrate-protein system with limited computational power, only the binding pocket and its immediate surrounding need to be solvated. A water sphere can be placed around the region of interest (binding pocket) and the rest of the system can be ignored. The water sphere can be divided into the reaction and reservoir region as shown in Figure 2.5. The dynamics of water molecules in the reservoir region are handled differently to the Newtonian manner by which the reaction region of the water sphere is treated. Langevin dynamics is used in the reservoir region to simulate the dynamical evolution of a system immersed in a larger system (larger water system). The explicit nature of the larger system is ignored. Two additional terms to Newton's second law is present in Langevin dynamics in order to emulate continuum effects.

$$F_{langevin}(r_i) = m_i \frac{\partial^2 r_i}{\partial t^2} = F_{internal}(r_i) - \xi \frac{\partial r_i}{\partial t} + F_{random}(t_i) \qquad (2.21)$$

In the above equation, ξ is the microscopic frictional coefficient and F_{random} is a random force generated to act on atoms in the reservoir region. The relationship between ξ and F_{random} regulates the temperature of the simulated system.[8, 19]

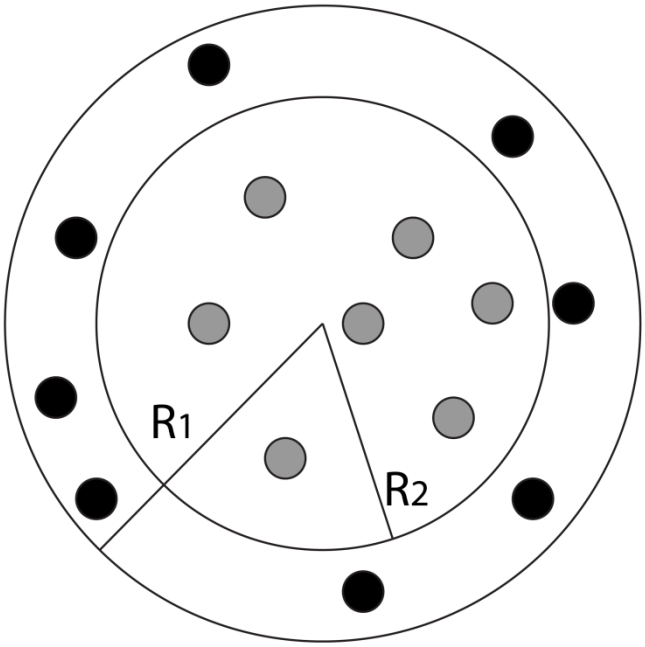


Figure 2.5 Division a system using stochastic boundary conditions. R_2 is the radius of the reaction region. $R_1 - R_2$ is the reservoir region.

2.4.1.3 Truncation Techniques

When accounting for nonbonded interactions, every pair of atoms in the system has to be accounted for. For a pair-wise model, the number of non-bonded terms increases in the order of approximately N^2, where N is the number of atoms. This can be extremely computationally intensive. Because the Lennard-Jones' potential falls off very rapidly, it is possible to introduce a cutoff distance. Any atom pair that has a distance beyond this cutoff value (r_C) will not be evaluated as a non-bonded interaction. This reduces the computational load of the simulation. When deciding the length of the cutoff value, the minimum image convention has to be taken into account and the size of the periodic box has to be considered. The cutoff should not be so large that a particle sees its own image.

Non-bonded neighbour lists can also be introduced to reduce the time taken to evaluate the number of non-bonded interactions. Using cutoff values alone still requires the distance between every pair of atom in the system to be calculated to identify whether or not they fall within the cutoff distance. Atoms within the cutoff distance do not change their position so much that it is necessary to

determine their distance relative to the cutoff for every time-step. Instead a non-bonded neighbour list can be used to store all atoms within the cutoff distance of each other. This neighbour list can be updated at specified times. It is important that the list is updates at a correct frequency. If the update frequency is too high, the procedure is inefficient and if the update frequency is too low, new atoms moving into the cutoff distance will be incorrectly handled. Energy discontinuities will be encountered when the update frequency is incorrect.

When introducing cutoffs, atoms near the cutoff distance experience a discontinuity in the potential energy and force near the cutoff distance. Energy conservation is required in molecular dynamics simulations. Shifting potentials and switching functions are the most commonly used techniques to deal with cutoff problems. For the shifting potential, a constant term is subtracted from the potential at all distance values *r* (Figure 2.6 (A)).[8]

$$U'(r) = U(r) - U \qquad r \leq r_c \tag{2.22}$$

$$U'(r) = 0 \qquad r > r_c \tag{2.23}$$

In the above equations, r_c is the cutoff distance and U_c is equal to the potential at the cutoff distance. Shifting potentials does however introduce a small discontinuity in the force because it suddenly drops the potential to zero for distances beyond the cutoff distance. This causes the potential to deviate from the true potential.

The switching function can be used to eliminate these discontinuities (Figure 2.6 (B)). When applying a switching function, a polynomial function (F_U) of the distance can either be multiplied over the entire Lennard-Jones potential, or between two distance values (r_D and r_C) along the Lennard-Jones potential (equations 2.24 – 2.26).[8, 24]

$$U'(r) = U(r) \qquad r < r_D \tag{2.24}$$

$$U(r) = F_U(r) \qquad r_D \leq r \leq r_C \tag{2.25}$$

$$U'(r) = 0 \qquad r > r_c \tag{2.26}$$

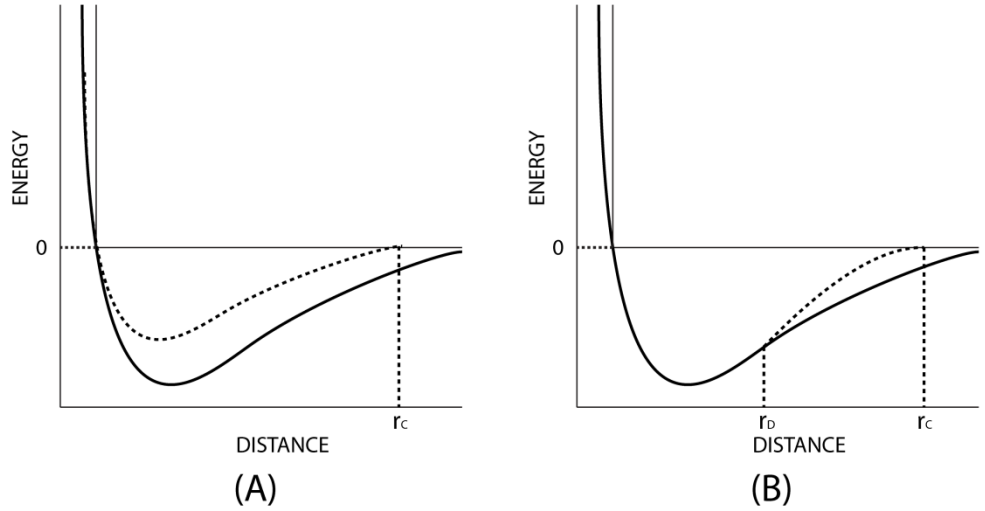


Figure 2.6 (A) Shifting potential. (B) Switching function between two distances r_D and r_C.

2.4.2 Water Solvation Models

Being able to accurately model the behaviour of the solvent system, which surrounds the biomolecular system being investigated, is of utmost importance. In enzymatic reactions, water molecules can be responsible for the stabilisation of the substrate in the binding pocket, or they can take part in the cleavage of bonds in hydrolytic reactions. The solvent also surrounds the entire protein and can induce structural changes to the surface as well as the core of the protein. The size of the solvent in the molecular system can vary from two water molecules to a few thousand water molecules. Large solvent systems pose computational complications due to simulation time limitations. Because of this problem, various solvation models have been formulated to address the needs of the system being investigated.[19]

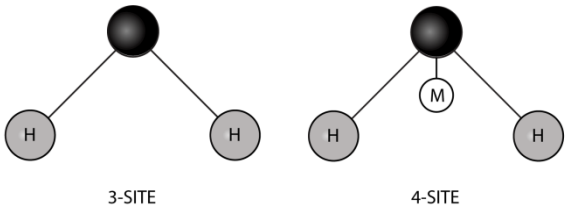
2.4.2.1 Implicit Water Solvation

Implicit or continuum solvation models allow for the removal of the evaluation of solvent-solvent interactions which could be computationally costly when simulating large condensed phase systems. In the implicit water model, the space that is occupied by the individual water molecules is modelled as a continuous medium, which possesses properties consistent with those of the solvent itself. Implicit water models facilitate modified conformational dynamics of protein structures, which can lead to irregular and misleading results.[9]

2.4.2.2 Explicit Water Models

In explicit water models, a water molecule possesses a definite physical presence. Representing waters explicitly is required when trying to accurately account for electrostatic and van der Waals interactions as well as to gain insight into specific solvent-solvent or solvent-solute interactions. In order to account for all these interactions, more computational time will be required. Water models referred to as "simple" models are most commonly used. These "simple" water models describe the water molecule as possessing a rigid geometry. "Simple" water models can possess anywhere between 3 to 6 interaction sites (Figure 2.7).

The TIP3P water model uses 3 interaction sites whereas the TIP4P water model uses 4 interaction sites.[25] In TIP3P a partial positive charge of 0.417 on each of the hydrogen atoms is balanced out by an appropriate negative charge of -0.834 on the oxygen atom. In the TIP4P model the negative charge (represented by *M* in Figure 2.7(B)) on the oxygen atom is shifted slightly along the bisector of the HOH angle towards the hydrogen atoms. This improves the electrostatic distribution around the water molecule. The van der Waals interactions are computed using a Lennard-Jones function with a single interaction point per molecule which is centred on the oxygen atom. No van der Waals interactions involving the hydrogen atoms are calculated.



	TIP3P	TIP4P
r(OH) (Å)	0.9572	0.9572
HOH (°)	104.52	104.52
q(O)	-0.834	0.0
q(H)	0.417	0.52
q(M)	0.0	-1.04
r(OM) (Å)	0.0	0.15

(A) **(B)**

Figure 2.7 (A) The TIP3P (3-site) and TIP4P (4-site) water models. Hydrogen atoms are grey and oxygen atoms are black. The *M* represents a dummy atom which possesses a negative charge. (B) Table comparing the parameters of the TIP3P and TIP4P models *r(OH)* and *r(OM)* are the distance between the O-H and O-M atoms respectively. *q(O), q(H)* and *q(M)* represent the charge on oxygen, hydrogen and the dummy atom respectively. HOH is the angle formed by the hydrogen and oxygen atoms.

There are other "simple" water models such as SPC, SPC/E, BF and ST2, to name a few. All these models are variations on a common theme. The number of interaction sites, the geometry of the water molecule and charge locations are varied in order to obtain a more accurate water model. Each model has positive and negative qualities. The SPC model better reproduces the structural and diffusion characteristics than the TIP3P and TIP4P models. TIP3P and TIP4P models better reproduce experimental results over a range of pressure and temperature values. The CHARMM force field was parameterised using the TIP3P water model and because of this, the TIP3P water model was used in this thesis for all simulations.[20]

2.5 Statistical Thermodynamics

Molecular dynamics is sufficient for the determination of the motion and trajectories of the atoms in a system. Information gained from these methods give insight into what is happening to the system at a microscopic level. In order to use microscopic information and convert it to macroscopic information such as the pressure, internal energy and Gibbs energy, statistical thermodynamics has to be implemented. These thermodynamic properties can be calculated from

the microscopic system using statistics. In the average molecular system, there are normally 10^{23} atoms. To solve the equations of motion for every single particle in the system will be impractical. To ease the burden of such large calculations, statistical mechanics does not enquire about the behaviour of individual atoms that are involved in the system, instead the average properties of the atoms in the system are calculated.[9]

2.5.1 The Ensemble

In thermodynamics, all that really matters are the bulk properties of the system. Bulk properties include the temperature *T*, pressure *p* and volume *V*, just to mention a few. An ensemble is a large number of representations of the molecular system. In molecular dynamics, every representation of the system at each time-step can be considered as replications of the original system where the momenta and positions of the atoms in the system change and evolve according to the restriction placed on them by the force field being used. However the ensemble conditions have to be maintained. In order for the number of representations *N*, to be considered collectively as an ensemble, certain conditions have to be maintained. There are 3 primary ensembles, namely; the canonical ensemble (NVT), the micro canonical ensemble (NVE) and the isothermal-isobaric ensemble (NTP). In order to obtain thermodynamic averages using the canonical ensemble, it is necessary to know the probability ρ, of finding the system in a given microstate in phase space. This probability is obtained using the Boltzmann distribution function $\rho(r,p)$.[9]

$$\rho(r,p) = \frac{e^{-H(r,p)/k_B T}}{Z} = \frac{N_i}{N_{t\ c}} \tag{2.27}$$

In the above equation, $H(r,p)$ is the Hamiltonian which is equal to the total energy of the system at the specified coordinates *r,* and momenta *p*. N_{total} and N_i are the total number of particles and the number of particles in state *i.*

$$H(r,p) = K(p) + U(r) = \sum_i \frac{p_i}{2m_i} + U(r) \quad (2.28)$$

The Hamiltonian is divided into the kinetic energy *K*, and the potential energy *U* where the potential energy is obtained from the force field being used for the simulation. The denominator in equation 2.27 above is the canonical partition function where,

$$Z(r,p) = \sum_i e^{\frac{-H(r,p)_i}{k_B T}} \quad (2.29)$$

and k_B is the Boltzmann factor. The canonical partition function can be used to calculate the internal energy, the Helmholtz free energy and the entropy. As the dynamics simulation continues and more regions of phase space are sampled, the distribution function takes on and converges to a more representative form of the molecular system. The distribution function can be used to calculate phase space *(q)* averages of any dynamic variable <*A*>.

$$< A(r,p) >_q = \int_V dr \int_{-\infty}^{\infty} dp p(r,p) A(r,p) \quad (2.30)$$

These are called thermodynamic or ensemble averages and they take into account every possible state of the system. Sampling every single value of phase space would be extremely difficult and is not very practical. Molecular dynamics allows for the sampling of regions of phase space that makes biomolecular sense. Averages calculated using dynamics are referred to as dynamic averages.[8, 19]

2.6 Protocol for Performing a Molecular Dynamics Simulation

For all the simulations carried out in this research, the following protocol was followed in order to ensure the quality of the results obtained.

1. Initial Preparation

   The initial starting coordinates for all atoms which make up the biomolecular system being simulated need to be acquired. The two most common methods for acquiring the starting coordinates are by the use of molecular modelling software packages or by directly obtaining the x-ray crystallographic structure of the protein or system. Sometimes it is necessary to use the molecular modelling software packages in order to add missing atoms or residues to an x-ray crystallographic structure that has a low resolution with many atoms missing.

2. Minimisation

   Structures obtained from x-ray crystallographic data or built by the use of molecular modeling software packages may not necessarily be in a realistic conformation and it is therefore necessary to determine more stable conformations of the system by the use of minimisation methods.

3. Heating

   After reducing steric clashing of atoms in the system, it is necessary to assign velocities to the atoms in the system. Starting the simulation of the system off at the desired temperature is not advisable as this can lead to unpredictable trajectories. Instead, the velocities are acquired from the Maxwell-Boltzmann distribution to simulate low temperatures initially for short durations. The temperature is gradually increased to the desired value.

4. Equilibration

   Once the system has reached the required temperature, certain conditions have to be placed on the system that will allow it to reach thermodynamic equilibrium. Initially the NPT ensemble is simulated to allow the volume of the system to evolve and the fluctuations in the temperature to minimise. The amount of time required for equilibration is dependent on the size of the system. A small system of approximately 1000 atoms might require 100 – 1000ps whereas larger systems of about

10000-20000 atoms might require 1 – 5ns of equilibration time or more. Root mean squared deviations (RMSD) of the simulated system can be used to identify whether or not the system has reached equilibrium.

5. Production

   After the system has reached thermodynamic equilibrium and fluctuations in the volume and temperature have been reduced, the actual dynamics can be performed. The equilibrated structure will be the initial starting conformation of the system in the production step. In this step, the time evolution of the system will be followed and all analyses will be carried out on data obtained in the production step.

2.7 Protein Preparation

There is always the possibility that structures obtained from x-ray crystallographic data have not been refined correctly. Due to limited resolution and imperfect phase information of the crystal structure, crystallographers have to sometimes rely on experience in order to complete the structure. Common errors that can be made when refining an x-ray crystallographic structure include overlooking residues in the structure as well as model-building errors which can lead to incorrect main-chain or side-chain conformations. Errors such as these are easy to make when dealing with x-ray resolutions of 2Å or lower. Building and refining a protein based on crystallographic data is not an exact science and it is therefore necessary to solve as many errors in the structure as possible before continuing to simulation analyses. There are several software programs capable of correcting the above mentioned problems and the WHATIF software packaged was used in this thesis.

2.7.1 Protein Structure Refinement

Protein verification "check" algorithms function by comparing the structural parameters taken from the incomplete crystal structure to standard structural parameters obtained from over 300 x-ray structures that have resolutions of

1.2Å or lower. Consider the length of a carbon-carbon bond (C-C). The measured C-C value from a incomplete x-ray structure would be compared to the normal distribution of C-C bond lengths obtained from reliable high resolution x-ray structures (Figure 2.8). Any value (x) that is more than 4 standard deviations away from the mean is considered an outlier and has to be investigated further.

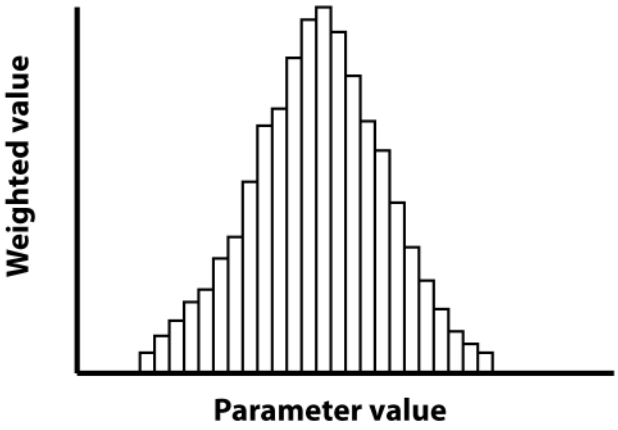


Figure 2.8 Normal distribution of parameter value being checked.

To determine how many standard deviations a value differs from the mean, the Z-score should be evaluated.

$$Z = \frac{x - \mu}{\sigma} \tag{2.31}$$

In equation 2.31, σ is the standard deviation and μ is the mean value of the normal distribution. Z is negative if the parameter being checked (x) is less than μ, and positive if x is greater than μ. If Z<-4 or Z>4 then the x value is considered to be an outlier and must be re-evaluated as to what kind of bond it is.

$$RMS_Z = \sqrt{\frac{\sum(Z^2)}{number\_of\_bonds}} \tag{2.32}$$

The "root mean squared Z-score" (RMS-Z) can be used to check the integrity of the x-ray structure. The RMS_Z value should approximately equal 1.0 for the

structure to be considered "good". Any x-ray structure which deviates from RMS-Z = 1 is considered problematic and indicates that there are many outliers present in the x-ray structure.[26]

2.7.2 Protein Protonation

It is often not possible to observe every atom in x-ray crystallographic structures. In some cases there are so much residual disorder in certain atoms that the resulting average electron density cannot be recorded. Hydrogen is a weakly scattering atom and is routinely invisible. In order to obtain the complete structure of a protein obtained from a crystal x-ray, the protonation states of titratable groups in the protein have to be determined under the specified pH conditions.

There are several methods used for the calculation of pK_a values from protein x-ray structures.[27, 28] A common issue that has to be addressed is whether the titratable groups in the proteins are in their correct protonation states. The protonation states of titratable groups in amino acids in a protein need to be determined and applied to the protein. This is of extreme importance because if a titratable group in the active site of the enzyme were to be protonated incorrectly, then it would fail to coordinate to the ligand and therefore fail to behave as it would during *in vivo* conditions. To determine the correct protonation states of a protein, the WHATIF software package was used. This software uses free energy methods to perform pK_a calculations. For acid-base reactions, the Henderson-Hasselbalch equation is used to determine the pK_a of a molecule.

$$pH = pK_a + \log [A^-]/[HA] = pK_a + \log K_G \qquad (2.33)$$

The situation is more complicated in a protein molecule because there are more species to consider. WHATIF splits up the effect of the protein environment. The free energy difference between the neutral and charged states of an isolated titratable group is evaluated. The free energy difference between the neutral and

charged states of that titratable group when it is part of the protein is also determined. These two free energies are compared and the pK_a value is determined.

To determine the pK_a of a titratable group in a protein environment in the aforementioned way, the calculation is divided into three steps. First, the desolvation energy associated with moving the neutral as well as the charged form of the group from the water to its position in the protein is determined. The interaction energy of the neutral and charged form of the residue with the permanent dipole of the protein is then calculated and the pair wise interaction energy between titratable groups is calculated. The information obtained from these calculations allows for the plotting of a titration curve for the individual titratable group. From this titration curve and the given physiological pH conditions, it is possible to determine the protonation state of the titratable group.[29]

2.8 Empirical Force Field Parameterisation

The accuracy of a molecular dynamics simulation is primarily dependent on the quality of the force field being used. The quality of the force field in turn is dependent on the method and target data used to optimise the parameters in the force field. When extending a force field, the same approaches and target data from the same sources should be followed and obtained in order to maintain consistency within the force field.[30, 31] Figure 2.9 lists sources of target data for common force field terms.

In Figure 2.10, the parameterisation procedure for the CHARMM force field is presented. In order to extend or add any molecules to this force field, the above steps have to be followed. All details mentioned below refer to the parameterisation of the CHARMM force field.

The first step involves the selection of appropriate model compounds. Adequate experimental data has to exist for the model compound. However, QM data can be used when experimental data is absent. For consistency and accuracy, a level of theory no lower than HF/6-31G(d) should be used. Model compounds can also be broken down into smaller molecules that can be linked together to form the desired model compound.

Term	Target data	Source
Internal	Geometries	QM, electron diffraction, microwave, crystal survey
Force	Vibrational	QM, IR,
External	Conformation properties	QM, IR, NMR, crystal survey
	Pure solvent properties	Vapour pressure, calorometry
	Crystal Properties	X-ray and neutron diffraction, vapour pressure, calorimetry
	Interaction energies	QM, microwave, mass spectrometry
Atomic charges (q)	Dipole moments	QM, dielectric permittivity, stark effect, microwave
	Electrostatic potentials	QM
	Interaction energies	QM, microwave, mass spectrometry
	Aqueoues solution	Calorimetry, volume variations

Figure 2.9 Types and sources of target data used in empirical force field optimisation procedures.[20]

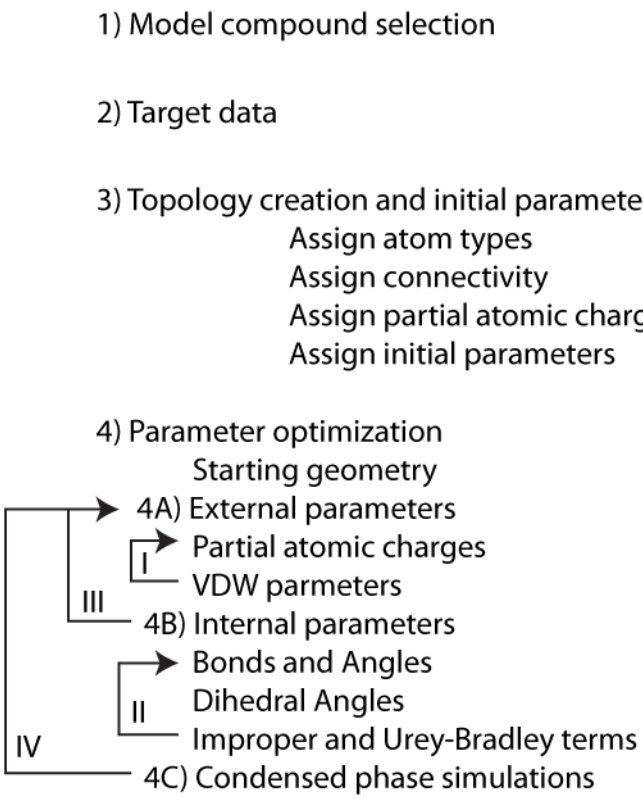


Figure 2.10 Steps involved in the preparation and extension of a force field. Included are iterative loops (I) over individual external terms, (II) over individual internal terms, (III) over external and internal terms and (IV) over the condensed phase simulations including both external and internal terms.[20]

The availability of target data can be a deciding factor in the choice of the model compound. The parameters of the model compound will be optimised to reproduce the selected target data. The availability of more data will allow for an increase in the accuracy of the parameters for the model compound. As mentioned before, the experimental data can be supplemented and extended with QM data. Using QM data alone can lead to inaccuracies. QM data is limited to the level of theory used in the force field being extended and is restricted to gas phase.

Topology information, which includes the connectivity, atom type and preliminary charges, must be set up after the model compound has been selected. This topology information can be obtained from molecules that have

similar conformation and configurations to that of the model compound. This information will be used as the initial parameters and will be adjusted as the parameterisation procedure continues.

The initial geometry or starting coordinates for the model compound can be obtained from x-ray crystal structures, or from QM determined methods. With the development of the Ewald method, parameterisation simulations can be performed in the condensed phase without the need for truncation methods. The parameter optimisation step involves several iterative loops. External parameters influence the final geometries and conformation energetics significantly, therefore in the parameterisation procedure, these parameters can be optimised first, followed by the internal parameters.[20, 31]

2.9 Simulation Analyses

Molecular dynamics simulations of proteins provide a wealth of information on an atomistic scale. The behaviour of a protein or solute with the solvent and substrate over time provides insight into structural changes as well as identifying key amino acids involved in the binding of the substrate. Water plays a pivotal role in protein functioning. By analysing interaction energies and distances of atoms or groups of atoms over time, useful information can be obtained, which will aid in the design of pharmaceutical molecules for various purposes.

2.9.1 Time Series

A time series refers to the configurations or conformations generated by a molecular dynamics simulation connected in time. A time series can be used to calculate time-dependent properties. Information regarding the interaction energies, distances, angles and dihedrals can be observed over time. This allows for the behaviour of the molecular system to be studied.

Correlation functions can also be used to determine whether or not two variables are dependent on each other over time.

$$C_{xy} = \frac{\frac{1}{M}\sum_{i=1}^{M} x_i y_i}{\sqrt{\left(\frac{1}{M}\sum_{i=1}^{M} x_i^2\right)\left(\frac{1}{M}\sum_{i=1}^{M} y_i^2\right)}} = \frac{<x_i y_i>}{\sqrt{<x_i^2><y_i^2>}} \quad (2.34)$$

In the above equation, there are M values of x_i and y_i time dependent data sets. If the above normalised function evaluates to 0, then no correlation exists, whereas if it evaluates to 1, a high degree of correlation exists. In biological systems, correlation functionscan be used to investigate the correlation between the geometric values of amino acids in a protein as it binds a substrate. Geometric data can include the bond lengths, angle and dihedrals of atoms in amino acids. Interaction energies can be correlated to substrate-amino acid distances. This is very useful for identifying important amino acids in a protein.

2.9.2 Hydrogen Bonding

Hydrogen bonds are central to the biological structure and function of protein folding and molecular recognition. Hydrogen bonds are the primary interaction involved in the binding of ligands to proteins. Electron densities for hydrogen-bonded protons are typically not observed in x-ray structures and thus bond formation must be inferred from the proximity of potential donor and acceptor groups. The ideal criterion to use in determining whether a hydrogen bond has been formed would be to measure the total interaction energy between the donor and acceptor groups. The interaction energy is a combination of electrostatics and van der Waals interactions in the hydrogen bond.[32]

Hydrogen bond analyses can also be based on geometric criteria. A distance of 2.4Å between the hydrogen bond donor and acceptor groups and an angle criteria of 120° is used. Hydrogen bond lifetimes of about 0.05-0.3ps are also taken into account. When considering protein hydrogen bond interactions, there are three common types. Hydrogen bonds can form between the protein and the

substrate, or between the protein and the solute, or through water-mediated interactions between the protein and the substrate (Figure 2.11).[33]

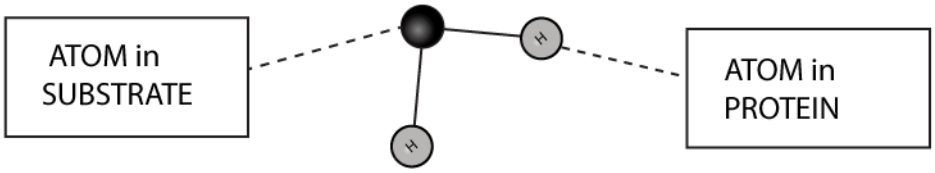


Figure 2.11 Water mediated interaction between a substrate and a protein.

2.9.3 Binding Pocket Volume

Being able to measure the changes in the volume of a binding pocket is very useful. Decisions on size and chemical structure of synthetic inhibitors can be decided upon based on the volume of a binding pocket. One method of measuring the volume of the binding pocket involves filling the binding pocket with hydrogen atoms. After the binding pocket has been filled, the volume of the individual hydrogen atoms is evaluated and summed together. This is a rough estimate of the volume of the binding pocket. This allows for relative comparisons to be made when comparing binding pocket volumes.[34]

2.10 Quantum Mechanics

Bound electrons in atoms do not follow classical mechanics and are limited to discrete energy levels. In order to evaluate systems based on their electronic structure, a different mechanics is required. De Broglie showed that matter can behave as a wave under certain conditions, and as a particle under different conditions. Quantum mechanics (QM) was developed to describe this dichotomy. There are various quantum theories for treating molecular systems. These quantum mechanical approaches are known as electronic structure approaches. The wavefunction (Ψ) forms a fundamental postulate in the theory of quantum mechanics.[8]

$$\theta\hat{\Psi} = e\Psi \tag{2.32}$$

Ψ which is an eigenfunction exists for any chemical system. θ is an operator which can act upon Ψ and return an observable property of the chemical system.

Ab initio and semi-empirical methods solve the wavefunction for the electronic Schrödinger equation. *Ab intio* methods use first principle methods whereas semi-empirical methods use experimental approximations to solve the wavefunction.

Density functional theory (DFT) also involves solving the Schrödinger equation, however the electron distribution is directly calculated using a density functional equation.

2.10.1 The Hamiltonian Operator

The operator in equation 2.32, which returns the energy *E*, of the system, as an eigenvalue, is referred to as the Hamiltonian operator *H*.

$$\hat{H}\Psi = E\Psi \tag{2.33}$$

The above equation is referred to as the Schrödinger equation. The time-independent Schrödinger equation that only depends on the spatial term or coordinates can be written as,

$$\left\{-\frac{\hat{\hbar}^2}{2m}\nabla^2 + V\right\}\Psi(r) = E\Psi(r) \tag{2.34}$$

where m is the mass, $\hbar$ is Plank's constant divided by 2π and the ∇^2 (del-squared) or Laplacian has the form,

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{2.35}$$

In equation 2.34, the Hamiltonian can be equated to,

$$\hat{H} = -\frac{\hbar^2}{2m}\nabla^2 + V \tag{2.36}$$

The Hamiltonian operator in equation 2.36 is composed of both kinetic and potential energy sections. The potential energy operator *V*, is described by,

$$V = \frac{1}{4\pi\varepsilon_0}\left(-\sum_i\sum_k\left(\frac{e^2 Z_k}{r_{ik}}\right) + \sum_{i<j}\frac{e^2}{r_{ij}} + \sum_{k<l}\frac{e^2 Z_k Z_l}{r_{kl}}\right) \tag{2.37}$$

where *i* and *j* run over the electrons, and *k* and *l* run over the nuclei. *e* is the charge on the electron, *Z* is the atomic number and *r* is the distance between electrons shown in subscript. The kinetic energy operator $\hat{T}$ is given by,

$$\hat{T} = -\frac{\hbar^2}{2m}\nabla^2 \tag{2.38}$$

Solving the Schrödinger equation for anything more complicated than the most simple of electron systems can be extremely difficult. Approximation have to be made in order to solve the wavefunction for large systems.

2.10.2 The Born-Oppenheimer Approximation

Equation 2.37 contains pairwise attraction and repulsion terms and no electron is moving independently of all the other electrons. In order to simplify this problem, the Born-Oppenheimer approximation can be used. Nuclei tend to move slower than electrons. Electron "relaxation" with respect to nuclear motion is instantaneous. Therefore, the electronic energies can be computed for fixed nuclear positions. After applying the Born-Oppenheimer approximation, the Schrödinger equation becomes,

$$\hat{H}_{elec}\psi_{elec} = E_{elec}\psi_{elec} \tag{2.39}$$

where the subscript 'el' refers to the electronic Hamiltonian. V_N is the nuclear-nuclear repulsion energy and q_i and q_k are the electronic and nuclear coordinates respectively. The semicolon indicates that q_i are independent variables, but q_k are the parameters of the wavefunction.

2.10.3 Molecular Orbital Theory

A wavefunction can be referred to as an orbital. Orbitals are 3-dimensional mathematical functions. A wavefunction for a polyelectron molecule can be represented as a linear combination of atomic orbitals (LCAO). This involves representing an arbitrary wavefunction for a molecule as a combination of more convenient atomic wavefunctions. These convenient wavefunctions are referred to as the "basis set".

$$\phi = \sum_{i=1}^{N} a_i \varphi_i \tag{2.40}$$

In the above equation, an estimate wavefunction ϕ, can be constructed from a linear combination of N atomic wavefunctions φ known as the "basis set". Basis sets define possible positions of electrons in space. Using larger basis sets leads to an increase in accuracy.[35, 36]

2.10.4 Hartree-product Wavefunctions

Removing the many electron problem and replacing it with a one electron Hamiltonian, makes the problem easier. The Hamiltonian can be written as,

$$H = \sum_{i=1}^{N} h_i \tag{2.41}$$

where the only terms are the one electron kinetic energy and nuclear attraction. N is the total number of electrons and h_i is the one electron Hamiltonian defined by,

$$h_i = -\frac{1}{2}\nabla_i^2 - \sum_{k=1}^{M}\frac{eZ_k}{r_{ik}} \qquad (2.42)$$

where M is the total number of nuclei. All eigenfunctions of the one electron Hamiltonian must satisfy the corresponding one-electron Schrödinger equation,

$$h_i\psi_i = \varepsilon_i\psi_i \qquad (2.43)$$

Due to the separable nature of a one electron Hamiltonian, the many electron eigenfunction can be constructed as products of one electron eigenfunctions.

$$\Psi_{HP} = \psi_1\psi_2 \ldots \psi_N \qquad (2.44)$$

Equation 2.44 is called the Hartree-product wavefunction Ψ_{HP}.[37]

2.10.5 The Hartree-Fock Self-consistent Field (SCF) Method

The Hartree-Fock SCF method is a basic molecular orbital calculation method. This method makes use of a Slater determinant wavefunction. All molecular wavefunctions are approximated. The SCF method replaces the many electron problem with a one electron problem. Electron-electron repulsions are treated in an average manner by making use of the SCF method in defining the Ψ_{HF} (equation 2.44) as a product of all individual electronic functions.

Convergence to the minimum energy is achieved through the use of the iterative SCF method. First an initial guess for spin orbitals is made. From this, the average field seen by an electron is calculated along with the corresponding eigenvalue equation. This equation is then used to calculate a new set of spin orbitals. This illustrates the self-consistent nature of the method. The process is

repeated until the spin orbitals from the previous integration are the same in the resulting eigenfunction.[8, 9, 39]

2.10.6 Density Functional Theory (DFT)

DFT was developed by Hohenberg and Kohn for the study of solids and extended by Kohn and Sham. DFT involves the determination of the energy as a function of the electron density whereas Hartree-Fock methods optimise a wavefunction. The total energy as a function of the electron density ($E_{DFT}[\rho]$) can be separated into several terms.

(2.49)

E^T and E^V represent the kinetic energy and nuclear-electron attraction energy respectively. E^J is the electron-electron repulsion term and E^X is the exchange correlation term, which includes the remaining electron-electron interactions.[40]

2.10.7 Semi-Empirical Methods

Hartree-Fock methods can be computationally expensive when dealing with large system sizes. Semi-empirical methods solve an approximate form of the Schrödinger equation. There are a number of semi-empirical methods available. Popular among these methods are MNDO, AM1 and PM3. Semi-empirical methods are computationally inexpensive compared to *ab initio* methods. Due to the experimentally parameterised nature of semi-empirical methods, a decrease in accuracy compared to *ab initio* methods is expected.

Reference:

1. J. A. McCammon, B. R. Gelin and M. Karplus, *Nature*, 1977, **267**, 585-590.
2. M. Karplus and G. A. Petsko, *Nature*, 1990, **347**, 631-639.
3. J. R. Knowles, *Phil. Trans. Roy Soc. Lond. B*, 1991, **332**, 115-121.
4. P. Koehl and M. Levitt, *J. Mol. Biol.*, 1999, **293**, 1161-1165.
5. C. Wong and J. A. McCammon, *A. Adv Protein Chem.*, 2003, **66**, 87-121.
6. G. Moraitakis, A. G. Purkiss and J. M. Goodfellow, *Rep. Prog. Phys.*, 2003, **66**, 383-406.
7. Y. Duan and P. A. Kollman, *Science*, 1998, **282**, 740-747.
8. A. R. Leach, *Molecular Modelling: Principles and Applications*, Prentice Hall, New York, 2001.
9. C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, 2nd edn., John Wiley and Sons Inc., 2004.
10. N. L. Burket and N. L. Allinger, *Molecular Mechanics*, Washington, DC, 1982.
11. A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry*, First edn., Dover Publication, London, 1996.
12. B. R. Brooks, R. E. Bruccoleri, O. B. D., D. J. States, S. Swaminathan and M. Karplus, *Journal of Computational Chemistry*, 1987, **4**, 187-217.
13. N. Foloppe and A. D. MacKerell, *Journal of Computational Chemistry*, 2000, **21**, 86-104.
14. S. E. Feller and A. D. MacKerell, *J. Phys. Chem B*, 2000, **104**, 7510-7515.
15. M. Kuttel, J. W. Brady and K. J. Naidoo, *Journal of Computational Chemistry*, 2002, **23**, 1236-1243.
16. J. E. Eksterowicz and K. N. Houk, *Chem. Rev.*, 1993, **93**, 2439.
17. V. Bakken and T. Helgaker, *Journal of Chemical Physics*, 2002, **117**, 9160-9174.
18. J. A. McCammon and S. C. Harvey, *Dynamics of proteins and nucleic acids.*, Cambridfe University Press, Cambridge, 1987.
19. A. Hinchliffe, *Molecular Modelling for Beginners*, 2nd edn., John Wiley and Sons Ltd., West Sussex, 2008.
20. O. M. Becker, A. D. MacKerell, B. Roux and M. Watanabe, *Computational Biochemistry and Biophysics*, Marcel Dekker, New York, 2001.
21. D. J. Beeman, *J. Comput Phys*, 1976, **20**, 130-139.
22. R. W. Hockney, *Methods Comput Phys*, 1970, **9**, 136-211.
23. J. Verlet, *Phys Rev*, 1967, **159**, 98-103.
24. W. D. Rogers, *Computational Chemistry Using the PC*, 3rd edn., John Wiley and Sons Inc., New Jersey, 2003.
25. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *Journal of Chemical Physics*, 1983, **79**, 926-935.
26. J. Kuszewski, A. M. Gronenborn and M. G. Clore, *Protein Science*, 2008, **5**, 1067-1080.
27. J. E. Nielsen and G. Vriend, *Proteins*, 2001, **43**, 403-412.
28. J. Antosiewicz, J. A. McCammon and M. K. Gilson, *J. Mol. Biol.*, 1994, **238**, 415-436.
29. J. E. Nielsen, *J. Mol. Graph. Model*, 2006, **25**, 691-699.

30. S. E. Feller, D. Yin, R. W. Pastor and A. D. MacKerell, *Biophys J.*, 1997, **73**, 2269-2279.
31. D. Yin and A. D. MacKerell, *Journal of Computational Chemistry*, 1998, **19**, 404-417.
32. K. Morokuma, *Accounts of Chemical Research*, 1977, **10**, 294-300.
33. U. Niesar, G. Clementi, G. R. Kneller and D. K. Bhattacharya, *Journal of Physical Chemistry*, 1990, **94**, 7949-7956.
34. J. D. Durrant, C. A. F. de Oliveira and J. A. McCammon, *Journal of Molecular Graphics and Modelling*, 2011, **29**, 773-776.
35. G. G. Hall, *Proceedings of the Royal Society*, 1951, **205**, 541-552.
36. C. C. J. Roothaan, *Reviews of Modern Physics*, 1951, **23**, 69-89.
37. V. Z. Fock, *Physik*, 1930, **61**, 126-148.
38. J. C. Slater, *Physical Review*, 1930, **35**, 210-211.
39. J. A. Pople and R. K. Nesbet, *Journal of Chemical Physics*, 1954, **22**, 571-572.
40. K. Raghavachari and J. B. Anderson, *Journal of Physical Chemistry*, 1996, **100**, 12960-12973.

Chapter 3

Free Energy Methods

3.1 Introduction

Being able to accurately calculate free energy differences in complex biochemical systems would allow for fast computer-aided drug design.[1, 2] In the modern drug design industry the process starts out with a large number of potential ligands, and it would be advantageous if the amount of time required to filter out the less favourable ligands could be reduced. Free energy calculations along with experimentally obtained data can also be used to assess the validity of a parameterised force field.[3, 4]

Free energy is a property that dictates most physical properties and being able to understand the free energy behaviour of a molecular system provides insight into the solvation, diffusion, binding, folding and many other properties, of the system. Only with the development of fast computer systems could Zwanzig's[5] description of macromolecular free energy calculations be applied to complex systems.

The primary limitation to the accuracy of the absolute free energy calculation for a given system is the ability to sample the total accessible phase space of the system. The two most commonly used techniques for sampling regions of phase space are molecular dynamics and Monte Carlo simulation techniques. However, several other computational approaches exist for the calculation of free energies, including continuum dielectric models and integral equation methods.[6, 7] All of these techniques suffer from insufficient sampling of the total accessible phase space of a system, and are therefore impractical for estimations of the absolute free energy of a system. A more practical approach would be to calculate the free energy difference between two closely related states. There are several methods currently in existence. In this chapter, Free Energy Perturbation (FEP),

Thermodynamic Integration (TI) and Slow Growth (SG) will be discussed and compared.[8]

Free energy is a state function and therefore does not depend on the manner in which a particular equilibrium state is reached. Using computer simulations, it is possible to modify or manipulate the energy function and allow for system transformations. These transformations could include alchemically "mutating" a residue and thereby convert a wild type protein into a mutant protein. In this thesis, free energy differences will be used to obtain relative free energies of binding of ligands to proteins. This will provide a potential route for lead improvements in drug design. Free energy can be expressed through the Helmholtz free energy A, obtained from the partition function Z,

$$A = -k_B T \, ln \sum_i e^{-H(r,p)_i/k_B T} = -\beta^{-1} \ln Z \qquad (3.1)$$

where β=(-k_BT)$^{-1}$ and k_B is the Boltzmann constant. Z represents the partition function mentioned in chapter 2. This equation provides a fundamental connection between thermodynamics and statistical mechanics.[8-10]

3.2 Free Energy Perturbation

In order to calculate the free energy difference ΔA, between two systems, or states of a system, the following equation can be used,

$$\Delta A = < A >_1 < A >_0 = -\beta^{-1} ln \frac{Z_1}{Z_0} = -\beta^{-1} ln < e^{-\beta \Delta H} >_A \qquad (3.2)$$

where Z_1 and Z_0 represent the canonical partition functions for state 1 (reference state) and state 0 (target state) respectively, and ΔH is the difference in the Hamiltonian energies. The quantity of interest between the two states of the system concerned is the excess Helmholtz free energy. When deciding on the reference state and the target state, the difference in these states should not be too large. Sampling the phase space of two very different states is difficult and

can lead to inaccurate results. In order to ensure good overlap of the phase space in the reference and target states of the system, non-physical intermediate states are used to connect them. This is referred to as the coupling parameter approach. The Hamiltonian will then be given as a function of the coupling parameter λ. Without loss of generality, we can choose 0≤λ≤1, such that λ=0 and λ=1 for the reference and target states of the system respectively. The Hamiltonian as a function of λ_i is evaluated using the linear function,

$$H(\lambda_i) = \lambda_i H_1 + (1-\lambda_i)H_0 = H_0 + \lambda_i \Delta H \tag{3.3}$$

where H_0 and H_1 denote the Hamiltonian of the reference and the target system respectively.[8] ΔH is the perturbation term in the target Hamiltonian, where $\Delta H = H_0 - H_1$. If N−2 intermediate states were chosen to link together the reference and target state such that λ_1=0 and λ_N=1, then the change in the Hamiltonian (ΔH) between two consecutive states is given by,

$$\Delta H_i = H(\lambda_{i+1}) - H(\lambda_i) = (\lambda_{i+1} - \lambda_i)\Delta H = \Delta\lambda_i \Delta H \tag{3.4}$$

where $\Delta\lambda_i = \lambda_{i+1} - \lambda_i$.[11] Given the above changes, the formula for the total free energy difference (equation 3.2) becomes,

$$\Delta A = -\beta^{-1} \sum_{i=1}^{N-1} \ln < e^{-\beta \Delta\lambda_i \Delta H(r,p)} >_{\lambda_i} \tag{3.5}$$

Since the Hamiltonian can be expressed as a function of λ, the canonical partition function can also be expressed as a function of λ.[8]

$$Z(r,p) = \sum_i e^{-H(r,p;\lambda)_i / k_B T} \tag{3.6}$$

3.3 Slow Growth

If the size of the λ intervals were to be reduced to a small enough window that the ΔH between any given interval ($\lambda_i - \lambda_{i+1}$) becomes arbitrarily close to zero, then it is possible to represent equation 3.5 as a truncated power series.

$$\Delta A = -\beta^{-1} lim_{d\lambda \to 0} \sum_{\lambda=0}^{1} ln\langle 1+\beta(H_{\lambda+d\lambda}-H_{\lambda})\rangle_{\lambda} \quad (3.7)$$

Since *ln(1+x)* can be approximated mathematically by *x*, for sufficiently small values of *x*, therefore,

$$\Delta A = -\beta^{-1} lim_{d\lambda \to 0} \sum_{\lambda=0}^{1} \langle \beta(H_{\lambda+d\lambda}-H_{\lambda})\rangle_{\lambda} \quad (3.8)$$

and further simplified to,

$$\Delta A = \lim_{d\lambda \to 0} \sum_{\lambda=0}^{1} (H_{\lambda+d\lambda} - H_{\lambda}) \quad (3.9)$$

Because the Hamiltonian is infinitesimally perturbed at every step in the simulation, the system will constantly be at a state of equilibrium, therefore ensemble averages are not required. That is why in equation 3.9 the ensemble average over λ is removed. This reflects the protocol of the Slow Growth technique. Every small change in λ is a step in the simulation as well, whereas in TI and FEP, multiple simulations over different windows of Δλ are involved.[10, 12]

3.4 Thermodynamic Integration

The Helmholtz free energy difference can be determined by a third simulation protocol known as Thermodynamic Integration (TI). Equation 3.9 can further be manipulated to,

$$\Delta A = \lim_{d\lambda \to 0} \sum_{\lambda=0}^{1} \langle (H_{\lambda+d\lambda} - H_{\lambda}) \rangle_{\lambda} \tag{3.10}$$

by summing over discreet intervals of Δλ,

$$\Delta A = \lim_{\Delta\lambda \to 0} \sum_{\lambda=0}^{1} \left\langle \frac{(H_{\lambda+d\lambda} - H_{\lambda})}{\Delta\lambda} \right\rangle_{\lambda} \Delta\lambda \tag{3.11}$$

then realising the relationship between the sum and a definite integral, and using partial derivatives.

$$\Delta A = \int_{\lambda=0}^{1} \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{\lambda} d\lambda \tag{3.12}$$

We can then reconvert to the summation process of approximate integration. The trapezoidal rule, an approximate technique for calculating definite integrals, can be used.

$$\Delta A = \sum_{\lambda=0}^{1} \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{\lambda} \Delta\lambda \tag{3.13}$$

The partial derivatives remain in the equation, because most simulations evaluate the energy using *H(λ)* functions like equation 3.3 that are trivially differentiated. Δλ in the final line is not considered to be small as in the SG technique, instead Δλ is a discreet interval that one would use when implementing the trapezoidal rule for definite integrals.[12]

FEP and TI are considered similar in that they involve multiple simulations over different windows of Δλ. In both techniques, accuracy in the method increases as the number of Δλ windows increase, and the size of the Δλ window decreases.[10]

3.5 Applications of Free Energy Calculations

The alchemical transformation is the process whereby the reference molecule is transformed into a target molecule via a pathway of nonphysical (alchemical) states. To aid in choosing the correct transformation pathway, thermodynamic cycles are used. These are hypothetical transformations that can only be carried out computationally and not experimentally. The thermodynamic cycle is reversible. Free energies associated with the forward or backward transformation can be calculated.[8]

The primary and most common use of these cycles is for the determination of the free energy of biological ligands in order to compare their binding affinities. Absolute free energy and relative free energy methods are commonly used in the calculation of free energies.

3.5.1 Relative Free Energy Calculations

Consider two ligands, I_1 and I_2, which could be inhibitors of an enzyme P. If ΔA_1 and ΔA_2 are the free energy of binding to the enzyme for inhibitors I_1 and I_2 respectively, then the relative binding affinity is $\Delta\Delta A = \Delta A_2 - \Delta A_1$. In order to simplify this calculation, we can consider using a thermodynamic cycle as shown in Figure 3.1. Because free energy is a state function, from Figure 3.1, $\Delta A_1 + \Delta A_4 = \Delta A_3 + \Delta A_2$. ΔA_3 corresponds to the free energy difference of the two ligands, I_1 and I_2 in solution, and ΔA_4 is the free energy difference of the two ligands, I_1 and I_2 in intermolecular (protein) complexes in solution.

By rearranging the equation, $\Delta A_4 - \Delta A_3 = \Delta A_2 - \Delta A_1$. Therefore, instead of computing $\Delta A_2 - \Delta A_1$, one can compute $\Delta A_4 - \Delta A_3$ to obtain $\Delta\Delta A$. This alchemical transformation ($\Delta A_4 - \Delta A_3$) is easier to compute than $\Delta A_2 - \Delta A_1$, because only the ligand has to be alchemically transformed, whereas in the latter, the entire protein would have to be created.[8]

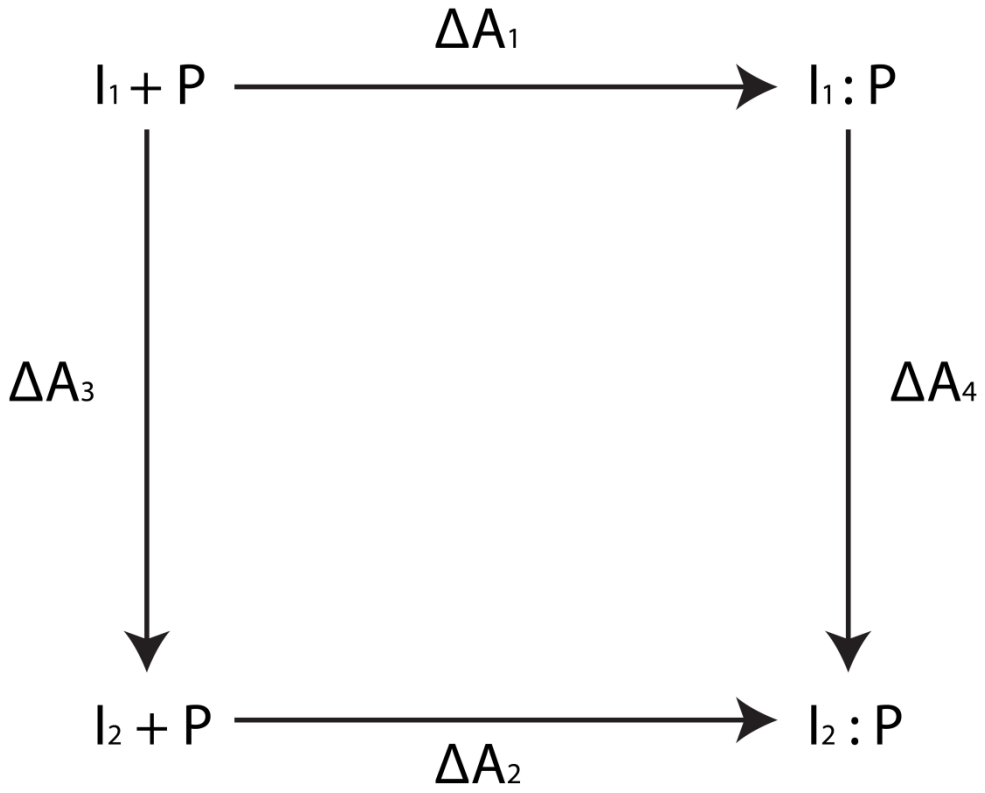


Figure 3.1 The relative free energy thermodynamic cycle comparing the binding affinity of I_1 and I_2 to the receptor enzyme P.[8]

3.5.2 Absolute Free Energy Calculations

The Absolute Free Energy change can be calculated using the "double-annihilation" method first proposed by Jorgensen et al (Figure 3.2). Instead of mutating a ligand into an alternate ligand and determining a relative binding affinity toward a common protein, in the annihilation method, the ligand is annihilated in the free (ΔA_3) and bound (ΔA_4) pathways of the cycle. The ligand is annihilated by cancelling the interaction of the ligand with its environment through the scaling of the nonbonded interactions, or by the scaling of the potential energy function. D in Figure 3.2 represents the annihilated ligand. From Figure 3.2, $\Delta A_2 = 0$ for D, and the absolute free energy is given by $\Delta A_1 = \Delta A_3 - \Delta A_4$.

As the protein-ligand interactions are reduced, the ligand may drift away from its original position as a result of a decrease in the attraction or repulsion forces from the protein. This leads to sampling problems. One way to prevent this from

happening is to introduce restraints in order to lock the ligand in place. This leads to loss of translational and rotational entropy, however there are analytical techniques that can make up for this loss. This method has been used in many applications, however, the relative free energy technique dominates in practice.[11]

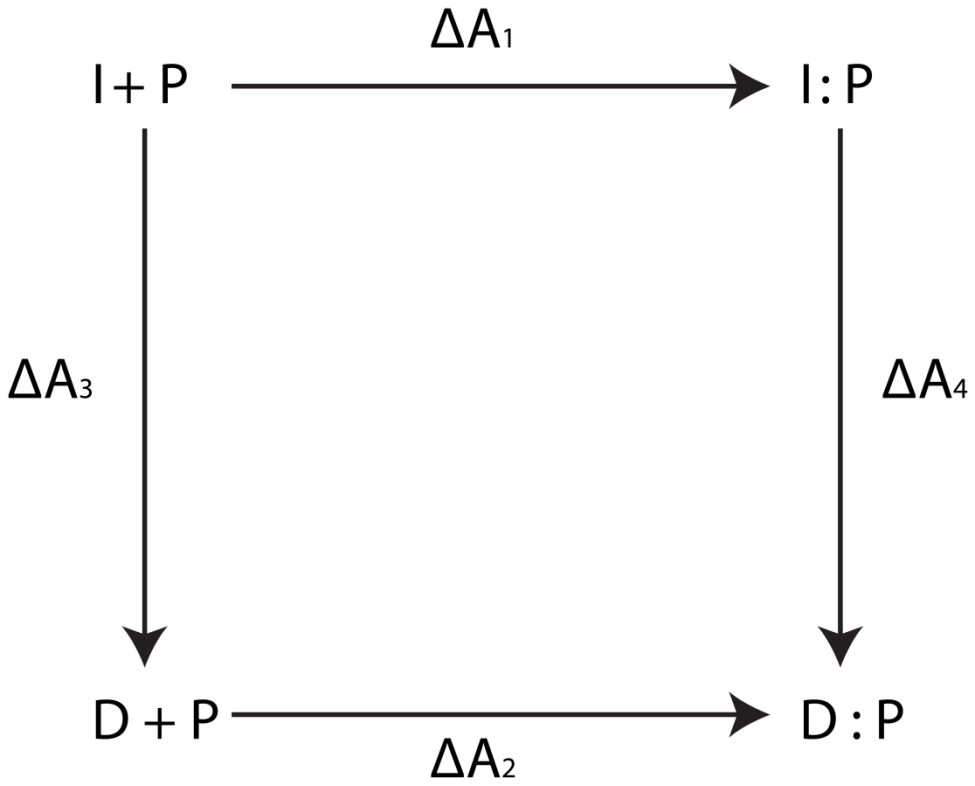


Figure 3.2 The absolute free energy thermodynamic cycle measuring the binding affinity of ligand I, to the receptor enzyme P. D represents a dummy ligand.[11]

3.6 Topological Paradigms

There are two primary approaches that are used to describe the topology of the reference state, target state and all intermediate states. They are referred to as the single-topology and dual-topology paradigms (Figure 3.3).

In the single-topology paradigm, both the reference and target topology are combined, whereby the most complex topology of the two states serves as the common denominator for both states. As the system transforms from the

reference state to the target state, the masses, charges, nonbonded parameters, bonds and angles are progressively changed as λ varies from 0 to 1. Dummy atoms are used to represent atoms that are not supposed to be present in a given state. Therefore, if the atoms do not form part of the target molecule, their bonds with other atoms progressively shrink to zero, and their point charges and van der Waals parameters are neglected as the transformation goes from the reference state to the target state, or going from λ=0 to λ=1.[8]

In the dual-topology paradigm, both the reference and target states coexist throughout the alchemical transformation. Using exclusion lists, atoms that are not common to both the reference and target states never interact in the simulation. As λ goes from 0 to 1, the intra and intermolecular interaction with the rest of the system are scaled by the λ value. Therefore when λ=0, only the interaction regarding the reference molecule is accounted for. And when λ=1, only the interaction regarding the target molecule is accounted for.

$$H(r;\lambda) = \lambda H_1(r) + (1-\lambda)H_0(r) \qquad (3.14)$$

The above equation shows the manner by which the Hamiltonians of the reference H_0, and the target states H_1, are scaled. Two problems are avoided when using the dual-topology over the single-topology paradigm. Firstly, the growing and shrinking of bonds is not required, and secondly the decoupling of electrostatic and non-electrostatic contributions during the simulation is no longer required.

Both topology paradigms suffer from a common problem known as the "end point catastrophe". At these end points the λ value approaches 0 or 1. Here, the unique reference or target state atoms are still interacting with the environment atoms. This interaction is weak, however the surrounding atoms are still able to clash against these appearing or disappearing atoms.[10]

In Figure 3.3, an alcohol group is alchemically transformed into a hydrogen atom. In the dual-topology method (Figure 3.3 (B)), the two black wedge functional

groups coexist but they do not "see" each other. In the reference state, the alcohol is expressed, but as λ approaches 1, the hydrogen atom becomes the expressed atom.

In the single-topology method (Figure 3.3 (A)), the oxygen atom of the alcohol group is mutated into a hydrogen atom. The hydrogen atom of the alcohol group is alchemically transformed into a dummy atom.

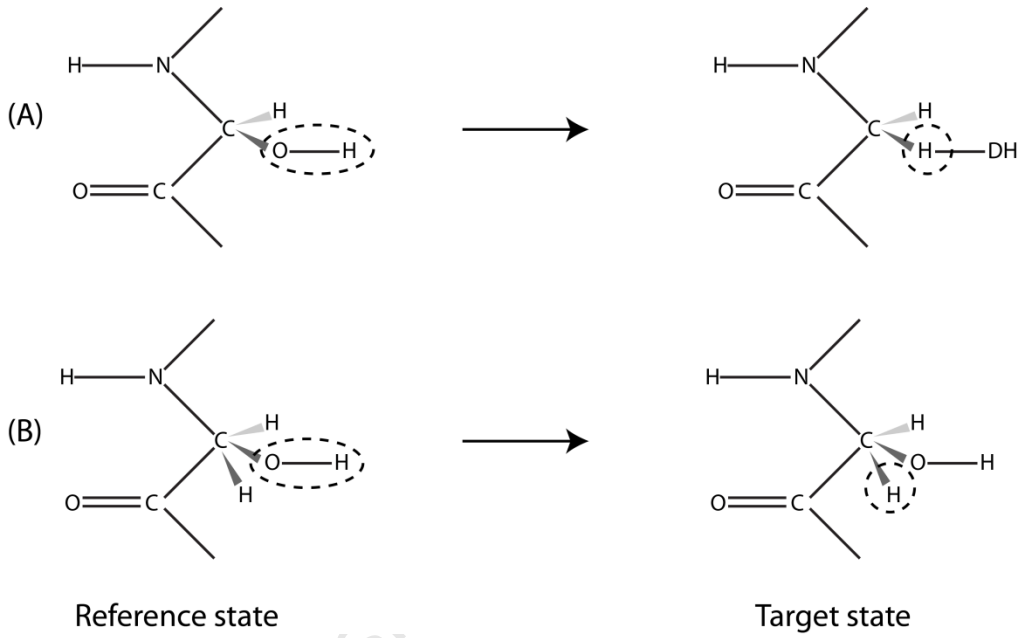


Figure 3.3 The (A) Single-Topology and (B) Dual-Topology paradigms illustrated as the molecule transforms from the reference state to the target state. Grey and black wedges represent bonds pointing into the page and bonds pointing out of the page. Dashed lines circle atoms relevant to the current state.[8]

3.7 Double-Wide Sampling

In double-wide sampling, the λ coordinate is divided into n sub-intervals. Consider the sub-interval where n=i in Figure 3.4. Performing a molecular simulation based on $\rho(r,p;\lambda)$ and using this dynamic data for calculating $\Delta H(\lambda_i \rightarrow \lambda_{i+1})$,

$$\Delta A(\lambda_i \rightarrow \lambda_{i+1}) = -\beta^{-1} \ln\left\langle e^{-\beta \Delta H(\lambda_i \rightarrow \lambda_{i+1})} \right\rangle_i \qquad (3.15)$$

which corresponds to the forward step from λ_i. Using the same dynamic data from the lambda dependent Boltzmann distribution function $\rho(r,p;\lambda)$, the $\Delta H(\lambda_i \rightarrow \lambda_{i-1})$ can be,

$$\Delta A(\lambda_i \rightarrow \lambda_{i-1}) = -\beta^{-1} \ln\left\langle e^{-\beta \Delta H(\lambda_i \rightarrow \lambda_{i-1})} \right\rangle_i \qquad (3.16)$$

which corresponds to the backward step from λ_i. This allows for the simultaneous forward and backward sampling from a single simulation.[8]

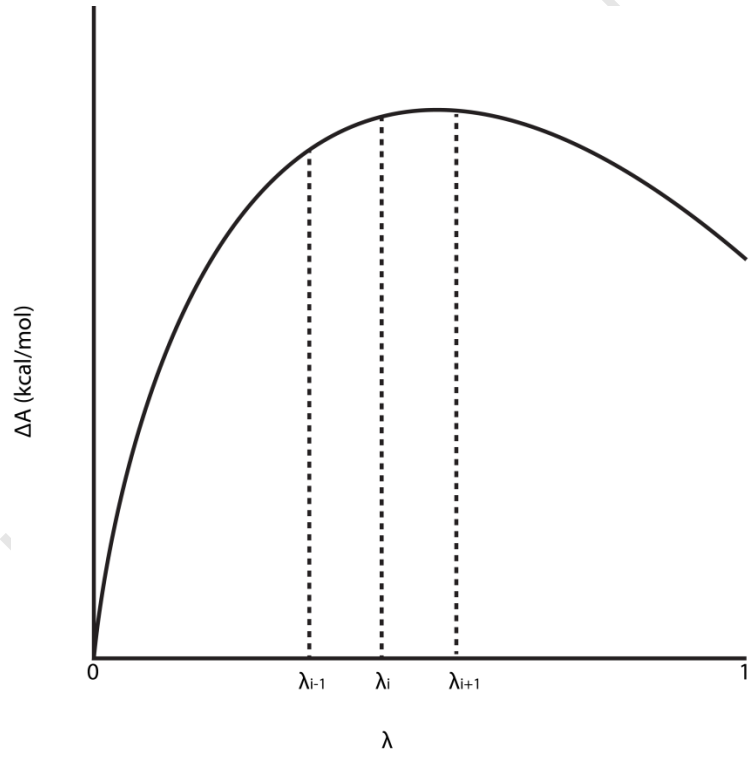


Figure 3.4 Double-wide sampling around the λ_i value.

3.8 Free Energy Calculation Protocol

The steps outlined below were followed in this research. These non-case-specific steps can be applied to any free energy method (FEP, TI or SG).

1. Setup topologies for the reference and target molecules.

2. Add force-field parameters for both states in the parameter file.

3. Realistic starting coordinates for the reference state have to be determined.

4. The protocol for the execution of a molecular simulation has to be completed (chapter 2).

5. A free energy method has to be decided on, be it FEP, TI or SG.

6. The number of λ values have to be decided upon, as well as the length of time required for equilibration and dynamics to be performed at each λ value.

3.9 Algorithm of the FEP Alchemical Transformation

The following steps presented form the algorithm used in this research for the determination of the ensemble average. The ensemble was generated using molecular dynamics and the dual-topology paradigm was preferred.

1. Build topologies and exclusion lists representing the reference and target states to prevent atoms not common in both states from interacting.

2. Generate ensemble conformations using molecular dynamics that represent the given λ value.

3. Evaluate the energy for each of the conformations generated for the reference state using the Hamiltonian $H(r,p;\lambda)$.

4. Repeat step 3 for the target state.

5. Evaluate the Hamiltonian energy difference (Equation 3.14).

6. Compute the ensemble average from which the free energy difference can be derived.

7. Increment λ value and repeat steps 2-6.

3.10 Reaching Convergence in Free Energy Calculations

Several issues can be encountered when implementing a free energy simulation. Most issues arise from trying to ensure optimal convergence as efficiently as possible.

In the master equation for both FEP and TI, the series of λ intermediates can be decided upon using various techniques. One such technique is to define a series of fixed width windows. If the λ window is too large, the Hamiltonian surfaces of $H(\lambda_{i+1})$ and $H(\lambda_i)$ will be too dissimilar, and therefore the required ensemble will converge slowly. The λ window can be reduced by increasing the number of λ points until the optimal λ window size is obtained.[11]

When using alchemical transformation in free energy simulations, the greatest convergence problems arise at the endpoint. The problem is due to the large qualitative change in the system on the first λ step in going from "nothing" and converting it into "something", or from going from "something" and converting it into "nothing". To reduce the effect of this problem, non-linear scaling[13] of the nonbonded interactions has been used as well as "bond shrinking"[14, 15].

Free energy calculations are highly dependent on the sampling of all phase space. Monte Carlo and molecular dynamics tend to sample only low energy regions along the pathway which connects the endpoints of the alchemical transformation.[16] Double-wide sampling, while being very efficient in obtaining two free energy differences from a single point, can add highly correlated error estimates, which can be unreliable when transformations have not been set up

correctly. Performing longer equilibration and dynamics phases during the free energy calculation will allow for sampling of more conformations in phase space and lead to an improved final result.[17]

Many aspects of the free energy calculation can affect the accuracy and efficiency of the end result. Selecting appropriate starting and ending structures in conjunction with the size and flexibility of the system are major factors that have to be considered.[11, 18]

Reference:

1. T. Simonson, G. Archontis and M. Karplus, *Acc. Chem. Res.*, 2002, **35**, 430-437.
2. C. Wong and J. A. McCammon, *J Am Chem Soc.*, 1986, **108**, 3830-3832.
3. A. Warshel, F. Sussman and G. King, *Biochemistry*, 1986, **25**, 8368-8372.
4. O. M. Becker, A. D. MacKerell, B. Roux and M. Watanabe, *Computational Biochemistry and Biophysics*, Marcel Dekker, New York, 2001.
5. R. W. Zwanzig, *J. Chem. Phys.*, 1954, **22** 1420-1426.
6. C. Brooks, M. Karplus and M. Pettitt, *Adv Chem. Phys.*, 1987, **71**, 1-259.
7. B. Roux and T. Simonson, *Implicit Solvent Model for Biomolecular Simulations*, Elsevier, Amsterdam, 1999.
8. C. Chipot and A. Pohorille, *Free Energy Calculations: Theory and Applications in Chemistry and Biology*, Springer, New York, 2007.
9. C. S. Tsai, *An Introduction to Computational Biochemistry*, Wiley-Liss Inc., New York, 2002.
10. C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, 2nd edn., John Wiley and Sons Inc., 2004.
11. H. Resat and M. Mezai, *Journal of Chemical Physics*, 1993, **99**, 6052-6061.
12. A. R. Leach, *Molecular Modelling: Principles and Applications*, Prentice Hall, New York, 2001.
13. A. J. Cross, *Chem Phys Lett*, 1986, **128**, 198-202.
14. D. A. Pearlman and P. A. Kollman, *J. Chem. Phys.* , 1991, **94**, 4532-4546.
15. L. Wang and J. Hermans, *J. Chem. Phys.*, 1994, **100**, 9129-9137.
16. A. E. Mark, S. P. van Helden, P. E. Smith, L. M. H. Janssen and W. F. van Gunsteren, *J Am Chem Soc.*, 1994, **116**, 6293-6299.
17. D. A. Pearlman and P. A. Kollman, in *Computer Simulations of Biomolecular Systems: theoretical and Experimental Applications*, eds. v. G. W. F. and P. K. Weiner, Escom Science Publishers, Netherlands, Editon edn., 1989.
18. M. Leitgeb, C. Schroder and S. Boresch, *Journal of Chemical Physics*, 2005, **122**, 084109-084115.

Chapter 4

Comparing Human and Herpes Uracil-DNA Glycosylase interaction with DNA

4.1 Introduction

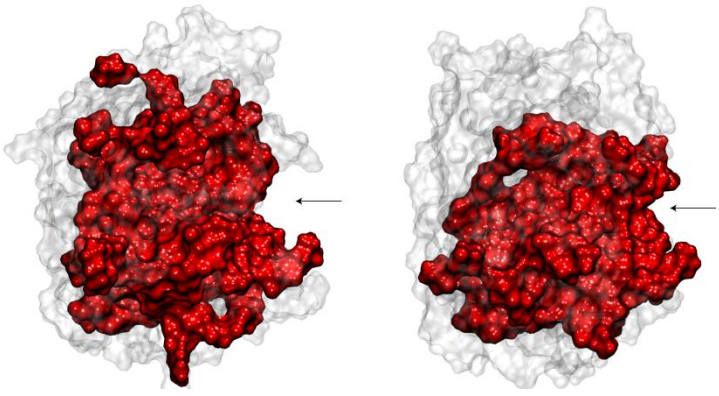
The presence of uracil in DNA results from one of two ways; either through the misincorporation of deoxyuridine triphosphate (dUTP) during cellular replication, or from the hydrolytic deamination of cytosine to produce a GUANINE:URACIL base pair mismatch.[1] If allowed to continue uncorrected, these mutations would lead to ADENINE:THYMINE transition mutations. The human and herpes simplex virus cell corrects this mutation via the base-excision repair pathway which is initiated by a uracil-DNA glycosylase (UDG) enzyme. Four families of the UDG enzyme have been identified. Family-1 enzymes are thought to adopt a "base sampling" method of identifying the uracil base.[2-4] For the base sampling method to occur, the uracil nucleotide has to assume the extrahelical conformation in order for the base to enter the binding site.[5, 6] Selectivity for uracil occurs through the favourable interactions between enzyme and base. The secondary structure and the amino acid composition and location of catalytic residues in the enzymes in family-1 UDG are highly conserved. It is known however that there are behavioural differences between these enzymes, which were mentioned in chapter 1.[7] In chapter 5, the differences in the inhibition character of these proteins are investigated.

In this chapter, two enzymes from the family-1 UDG are investigated. These are the herpes simplex virus 1 uracil-DNA glycosylase (hsvUDG) and the human uracil-DNA glycosylase (hUDG) enzymes. Two molecular dynamics simulations were carried out with DNA as a substrate in the binding pocket of the hUDG and hsvUDG enzymes. Important interactions between the DNA and the enzymes were then determined and rationalised.

4.2 Sequence and Structural Comparison between hsvUDG and hUDG

The common DNA repair function of these uracil-DNA glycosylase enzymes has resulted in convergent evolutionary similarities in their amino acid sequence and structure.[2] In Figure 4.1, it can be seen that the majority of the conserved amino acids are located in the regions including and surrounding the binding pockets. This illustrates the evolution of a conserved binding pocket for a common purpose, which is the recognition of the uracil base and the cleavage of the glycosydic bond between the deoxyribose sugar and the uracil base. Figure 4.2 reveals how identical the two proteins are. Computational analyses using alignment algorithms[8] were used to illustrate the similarities in the amino acid sequence (Figure 4.2). Amino acids are either termed identical when they possess the same location and molecular structure, or similar if they possess the same location and similar but not identical molecular properties.

Of the 232 amino acid sequence comparison, 93 amino acids are conserved. There is a 40.1% similarity in the amino acid sequence between the two proteins. In Figure 4.3, even though there is only a 40.1% similarity between the proteins, the secondary structures of the proteins also reveal large structural similarities with the α helix appearing to be the dominant structural motif conserved in both proteins.



(A) **(B)**

Figure 4.1 3-dimensional amino acid comparison between (A) hUDG and (B) hsvUDG. Identical amino acids are shown in red and the rest is shown in transparent grey. Arrows indicate binding pockets.

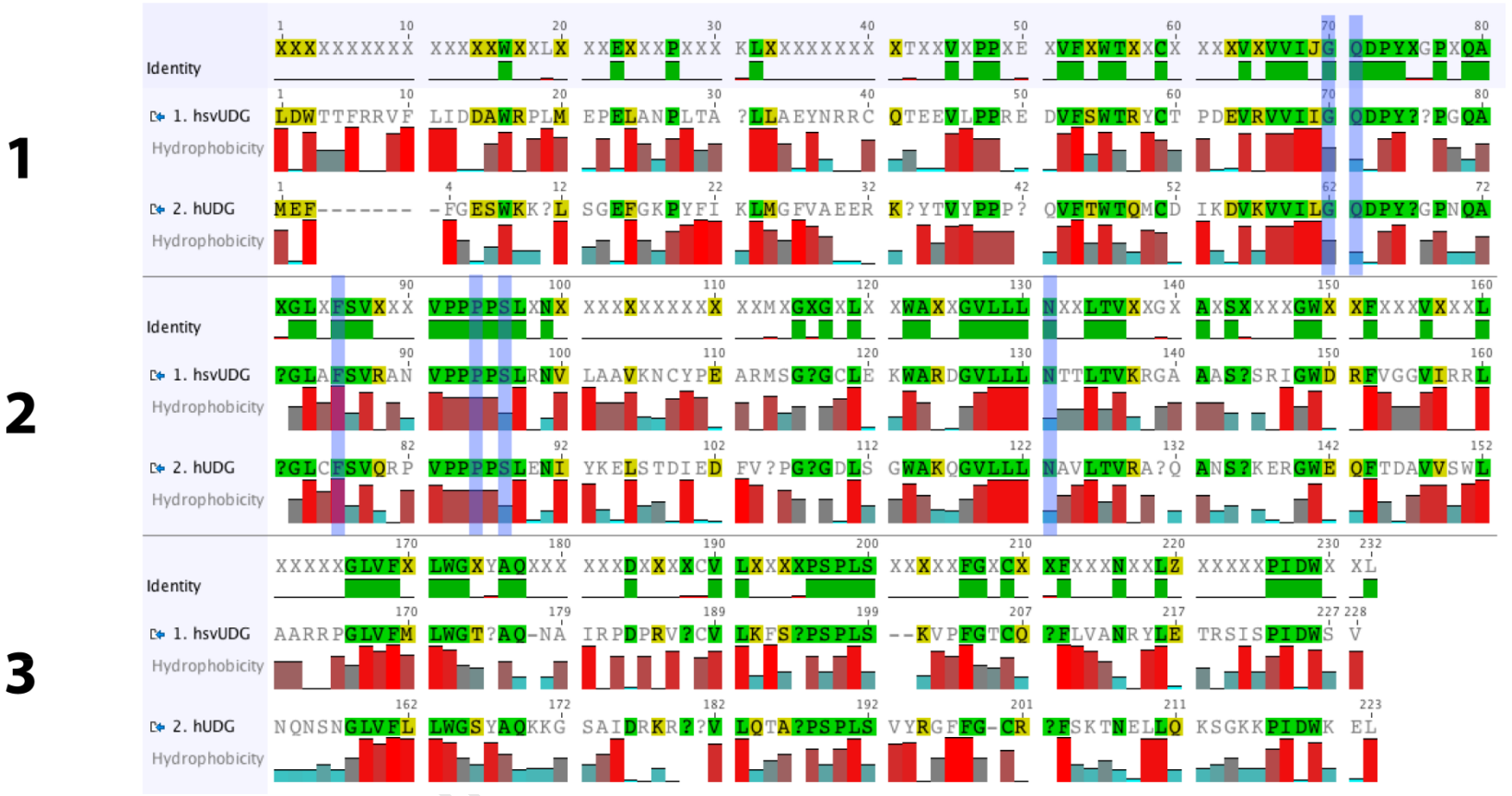


Figure 4.2 The sequence alignment and comparison of hUDG and hsvUDG. Green, yellow and white highlighted amino acids represent identical, similar and not similar respectively. Blue highlighted columns represent amino acids present in the binding pocket.

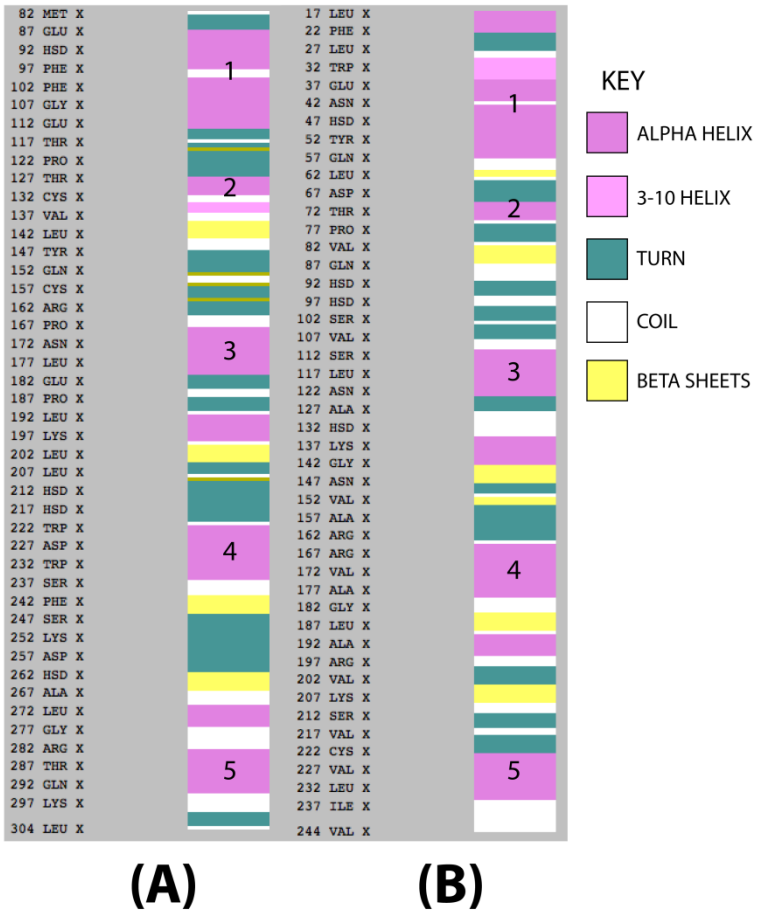


Figure 4.3 Secondary structure sequence comparison between (A) hUDG and (B) hsvUDG. Similar regions are numbered.

4.3 Methodology

4.3.1 Molecular Dynamic Simulations

To gain insight into the similarities and differences in the functioning of hUDG and hsvUDG, molecular dynamics simulations were performed on both enzymes. A DNA double helix consisting of 5 base pairs of which the extrahelical uracil base is included, was used as the substrate for both enzymes. The simulations were carried out using the CHARMM33b2[9] program employing the empirical energy function mentioned in chapter 2 (equation 2.1). The proteins were modelled using the CHARMM27[10, 11] all-atom force field which was designed to simulate proteins and nucleic acids. To solvate the systems, a truncated octahedron of length 70Å, originally containing 12161 TIP3P[12] water molecules, in order to obtain a water density of 1.0g.cm^{-1}, was used in both simulations. Water molecules with heavy atom distances within 3Å of the solute were then removed. The van der Waals nonbonded interactions were truncated using the switching function. The switching function was initiated at a distance of 10Å and truncated at 12Å from the atom concerned. All hydrogen bond lengths were kept constant using the SHAKE[13] algorithm. The electrostatic interactions were evaluated using the Ewald summation method[14, 15]. The nonbonded interaction list and solvent image were updated every 10fs. The systems were first heated gradually from 145K to 300K and then equilibrated for 8ns at a pressure of 1bar and a temperature of 300K using the isothermal-isobaric ensemble (NPT) to allow the system's volume to proliferate. This was followed by 7ns of production simulation using the canonical ensemble (NVT). Data for both simulations were stored at 10ps intervals. The standard deviation in temperature fluctuation during the production period is ±1.7K.

4.3.2 Preparation

Initial coordinates for both systems were obtained from the Brookhaven Protein Data Bank. The crystal structures used were resolved at 1.9Å and 1.75Å for 1SSP[16] (hUDG) and 1UDG[17] (hsvUDG) respectively.

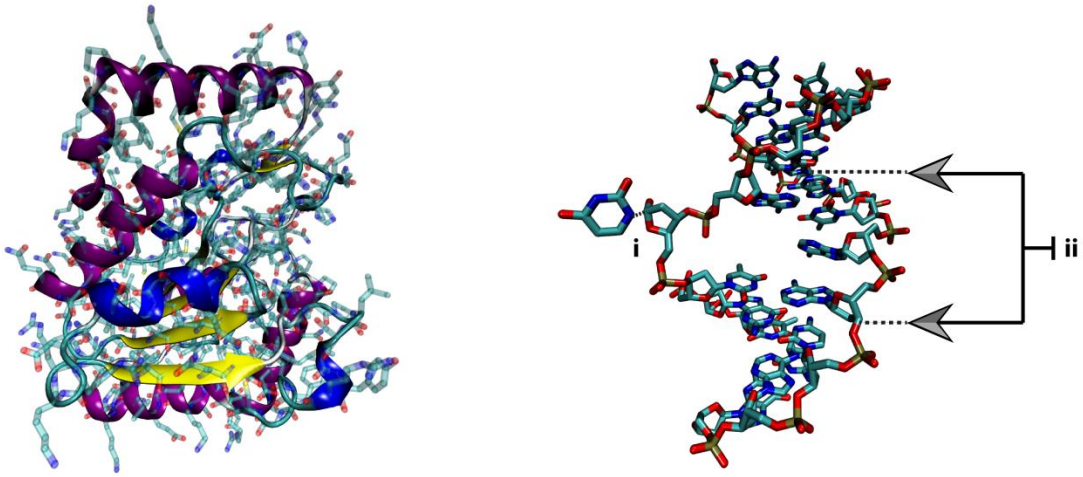


Figure 4.4 (A) Deprotonated protein structure and (B) DNA double helix obtained from Brookhaven Protein Data Bank (1SSP). (i) Cleaved glycosydic bond. (ii) Adjusted length of DNA used in simulations.

Incomplete side chains and protonation states of the crystal structures were checked and corrected to a pH of 7 using pK_a calculations[18] mentioned in chapter 2. Figure 4.4(A) shows the unprotonated protein obtained from the Protein Data Bank and Figure 4.4(B) shows the DNA obtained initially where the dashed lines at (i) indicate the glycosidic bond that was recreated and (ii) indicates the truncation of the DNA to include 5 base pairs. The overall charge of the protein structures after evaluating the protonation states of all titratable groups was determined to be +8. When evaluating the electrostatic nonbonded interactions using the Ewald summation method, it is good practice to ensure that the overall charge of the system is equal to zero.[15] This ensures efficient convergence of the Ewald sum. The 5 base pairs carry a negative charge of -8. The base sequence in the double helix is,

5'-D(*GP*TP*UP*AP*T)-3'	STRAND-1
5'-D(*AP*TP*AP*AP*C)-3'	STRAND-2

Once the protein and the DNA are combined to form a complex, the overall charge of the system then becomes zero (Figure 4.5). To obtain the starting

structure, the extrahelical uracil nucleotide was then docked inside the binding pocket.

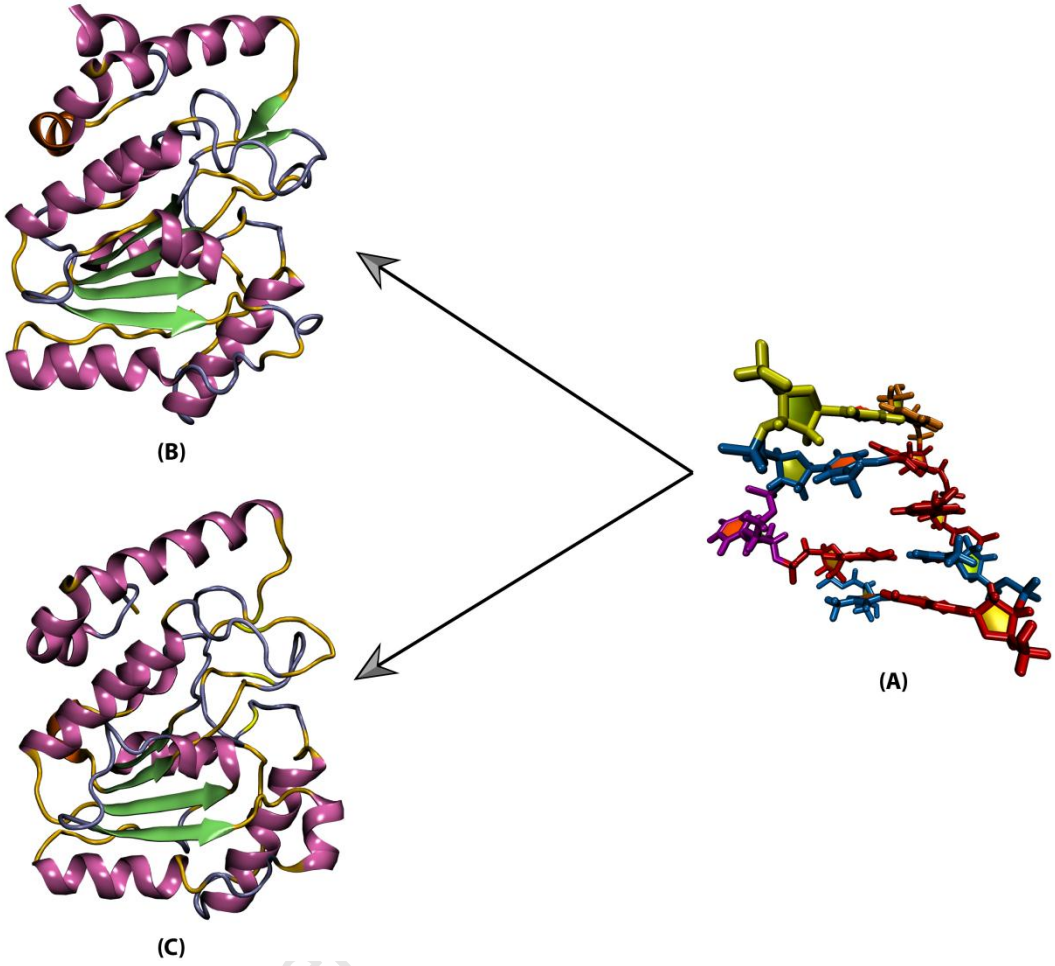


Figure 4.5 (A) 5 base pair DNA molecule used in the simulation of (B) hsvUDG and (C) hUDG.

4.4 Results and Discussion

4.4.1 Substrate – Protein Interactions

Proteins that have evolved over a long period of time are known to be highly substrate specific.[19] This specificity results from the unique and precisely placed amino acids in the protein binding pocket, which will have distinct interactions with a substrate of interest, once the substrate has entered the binding pocket.[20] In this work, interactions that will be investigated include hydrogen bonding[21], van der Waals, electrostatic, water-mediated[22] and π–stacking[23] interactions. Given that DNA is involved in these simulations, there will be numerous interaction between the DNA and the protein backbone, which will be primarily composed of van der Waals and electrostatic interactions.[22] Van der Waals forces include attraction between atoms, molecules and surfaces caused by correlations in the fluctuating polarisations of nearby particles. Hydrogen and water-mediated bonds are less common and mostly occur within the binding pocket between catalytic residues and the DNA substrate. Aromatic interactions such as π-stacking interactions are non-covalent interactions and are caused by overlapping of the π-orbitals of conjugated systems.

Due to the approximate nature of molecular dynamics, electron behaviour is ignored.[9] Therefore, in order for interaction to be classified accordingly in molecular dynamics simulations, specific geometric criteria have to be observed. The interaction energy referred to in this thesis is obtained from the electrostatic and van der Waals interactions, which constitute the nonbonded interaction set of the empirical energy function. Hydrogen bonds are defined as possessing a donor-hydrogen-acceptor angle of between 180°–120° and a hydrogen-acceptor distance of less than 2.4Å.[25] In order for effective π-stacking to be observed, the distance between participating π-systems is usually reported to be 3.00Å – 4.00Å.

4.4.2 Uracil nucleotide binding in hUDG

A combination of water-mediated bonds and direct hydrogen bonds stabilises the uracil base in the binding pocket producing an interaction energy of -33.517 kcal.mol^{-1} between the protein and the uracil base (Table 4.1). Figure 4.6 illustrates a 2-dimensional interaction profile between the hUDG enzyme and the uracil nucleotide. The amino acids in Table 4.1 are arranged according to various regions of interaction with the DNA. Amino acids in regions A, B and C are located near the extrahelical uracil base, the deoxyribose sugar of the uracil nucleotide and the rest of the DNA respectively.

There are several important interactions between the hUDG enzyme and the uracil nucleotide (Figure 4.6). The carboxamide functional group asparagine (ASN204) forms a bidentate interaction with the H3 hydrogen and the O4 oxygen of the uracil pyrimidine. Further, these interactions are thought to be responsible for distinguishing between cytosine and uracil.[24] The O2 oxygen of the bound uracil receives a hydrogen bond interaction from the NH of the peptide linkage between the conserved glycine (GLY143) and glutamine (GLN144) amino acids. The NH of the peptide bond between phenylalanine (PHE158) and cysteine (CYS157) forms a hydrogen bond interaction to the O4 oxygen of uracil. Tyrosine (TYR147) forms an H-π interaction with the H5 hydrogen of uracil. This interaction seems to be responsible for the selection of uracil over thymine.[24] Thymine has a bulky methyl group bonded to C5 of the pyrimidine whereas uracil consists of a hydrogen atom bonded to the C5 of the pyrimidine base. The methyl group would cause a greater degree of steric hindrance in the binding pocket and be very unfavourable. Histidine located at position 268 (HIS268) forms a water-mediated interaction with the O2 oxygen of uracil. Histidine located at position 148 (HIS148) forms a water-mediated interaction with the 3' oxygen of the deoxyribose sugar.

REGION	AA	Average Interaction Energy (kcal.mol^{-1})											
		Uracil Nucleotide				Uracil Base				Complete DNA Substrate			
		DISTANCE	VDW	ELEC	TE	DISTANCE	VDW	ELEC	TE	DISTANCE	VDW	ELEC	TE
A	ASP145	6.27	-1.90	-7.91	-9.81	5.50	-1.15	-2.86	-4.01	13.67	-2.22	-11.69	-13.91
	PRO146	7.48	-1.76	0.73	-1.03	5.43	-1.53	-0.79	-2.31	16.08	-1.93	-2.35	-4.28
	CYS157	8.41	-1.27	-1.97	-3.24	5.57	-1.22	-1.08	-2.30	19.97	-1.27	-2.12	-3.39
	HIS268	6.14	-2.67	-1.98	-4.65	5.43	-2.03	-0.23	-2.26	13.23	-5.43	-11.50	-16.93
	PHE158	5.86	-4.06	-4.05	-8.11	4.30	-3.19	-2.66	-5.86	17.31	-4.14	-4.61	-8.75
	ASN204	8.86	-0.76	-7.39	-8.15	6.15	-0.69	-8.58	-9.27	19.24	-0.76	-7.59	-8.36
	TOTAL A	**-**	**-12.42**	**-22.57**	**-34.99**	**-**	**-9.82**	**-16.19**	**-26.01**	**-**	**-15.75**	**-39.87**	**-55.62**
B	SER169	5.59	-0.13	-21.34	-21.47	7.10	-0.37	-1.61	-1.98	13.60	-0.52	-29.96	-30.49
	TYR147	6.99	-5.00	-0.45	-5.44	4.73	-3.10	-1.63	-4.73	13.92	-8.06	-10.41	-18.46
	GLN144	5.80	-2.50	-6.63	-9.13	5.77	-0.89	-2.68	-3.57	13.54	-4.67	-2.70	-7.37
	TOTAL B	**-**	**-7.63**	**28.42**	**-36.04**	**-**	**-4.36**	**-5.92**	**-10.28**	**-**	**-13.25**	**-43.07**	**-56.32**
C	SER270	6.78	-1.73	-10.20	-11.93	9.40	-0.06	-0.05	-0.11	10.93	-2.03	-23.13	-25.16
	PRO271	9.80	-1.06	-4.32	-5.38	12.52	0	0	0	10.55	-1.85	-7.89	-9.74
	LEU272	12.12	-0.08	-2.10	-2.19	15.12	0	0	0	8.76	-2.22	-7.51	-9.73
	HIS148	7.11	-1.62	-5.92	-7.54	7.72	-0.44	-1.40	-1.84	10.54	-6.40	6.30	-0.10
	PRO168	8.69	-2.14	-4.10	-6.23	10.62	-0.03	-0.20	-0.23	12.43	-3.55	-10.11	-13.66
	ALA214	8.00	-0.41	-3.27	-3.68	9.32	-0.05	-0.13	-0.26	8.77	-3.61	-6.81	-10.42
	ASN215	9.76	-0.14	-0.03	-0.17	10.57	-0.03	-0.28	-0.31	11.68	-3.32	2.58	-0.74
	GLN152	10.54	-0.11	-0.44	-0.55	10.05	-0.05	-0.63	-0.68	16.34	-2.34	-9.65	-11.99
	GLN213	12.14	-0.04	-0.58	-0.62	13.00	0	-0.19	-0.19	-3.61	-1.41	-2.20	-3.61
	HIS212	12.55	-0.02	-0.34	-0.36	13.74	0	0.13	-0.13	9.60	-3.28	-13.08	-16.36
	PRO167	-3.74	-1.07	-2.68	-3.75	8.81	-0.10	-0.46	-0.56	14.75	-2.21	-9.00	-11.21
	TYR248	10.01	-0.18	-1.54	-1.71	10.37	-0.07	-0.70	-0.77	15.45	-0.47	-16.06	-16.53
	TOTAL C	**-**	**-8.28**	**-35.83**	**-44.11**	**-**	**-0.81**	**-4.27**	**-5.08**	**-**	**-32.68**	**-96.57**	**-129.25**
TOTAL PROTEIN		**11.04**	**-32.15**	**-72.95**	**-105.10**	**8.32**	**-16.36**	**-17.16**	**-33.51**	**22.43**	**-66.69**	**-211.68**	**-279.37**

Table 4.1 Average interaction energies between the DNA substrate and key amino acid residues in the hUDG enzyme. Three regions of the substrate are considered. These regions include the uracil base, the uracil nucleotide and the complete DNA substrate. Amino acids in region A, B and C are located near the extrahelical uracil base, the deoxyribose sugar of the uracil nucleotide and the rest of the DNA respectively.

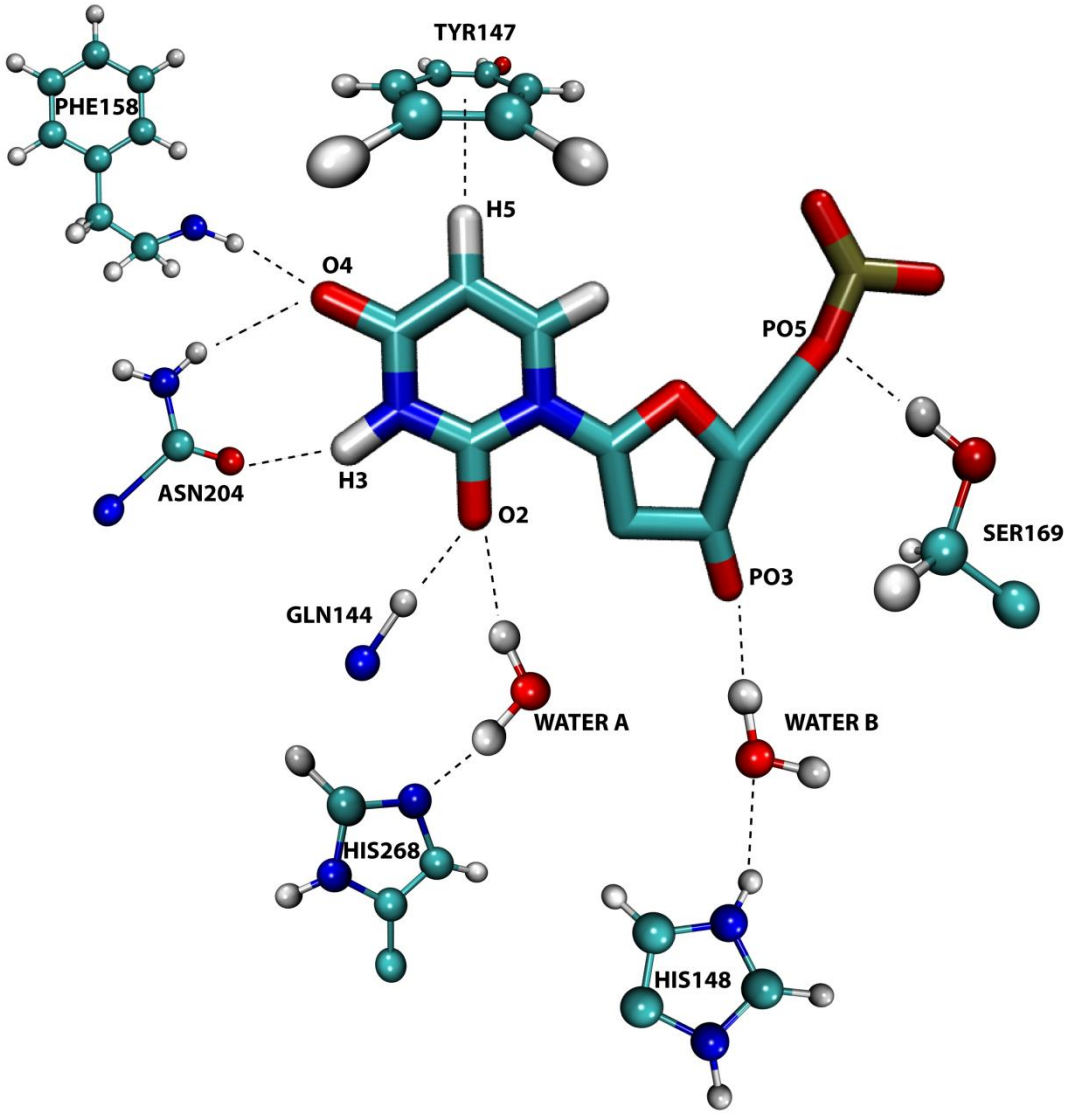


Figure 4.6 2-dimensional illustration of significant interactions between amino acids in the binding pocket of hUDG and the uracil nucleotide.

Figure 4.7 shows the time series graphs of the direct interactions shown in Figure 4.6, between the hUDG enzyme and the uracil nucleotide. Figure 4.8 and 4.9 show the time series graphs for the water-mediated interaction of HIS268 and HIS148 with the uracil nucleotide. A correlation in the 2 water molecules interaction with the uracil nucleotide and the histidine amino acids can be seen in (A) and (B) of Figure 4.8 and 4.9. This drop and rise in interaction energy is caused by the location of the water molecule as it moves towards and away from the substrate.

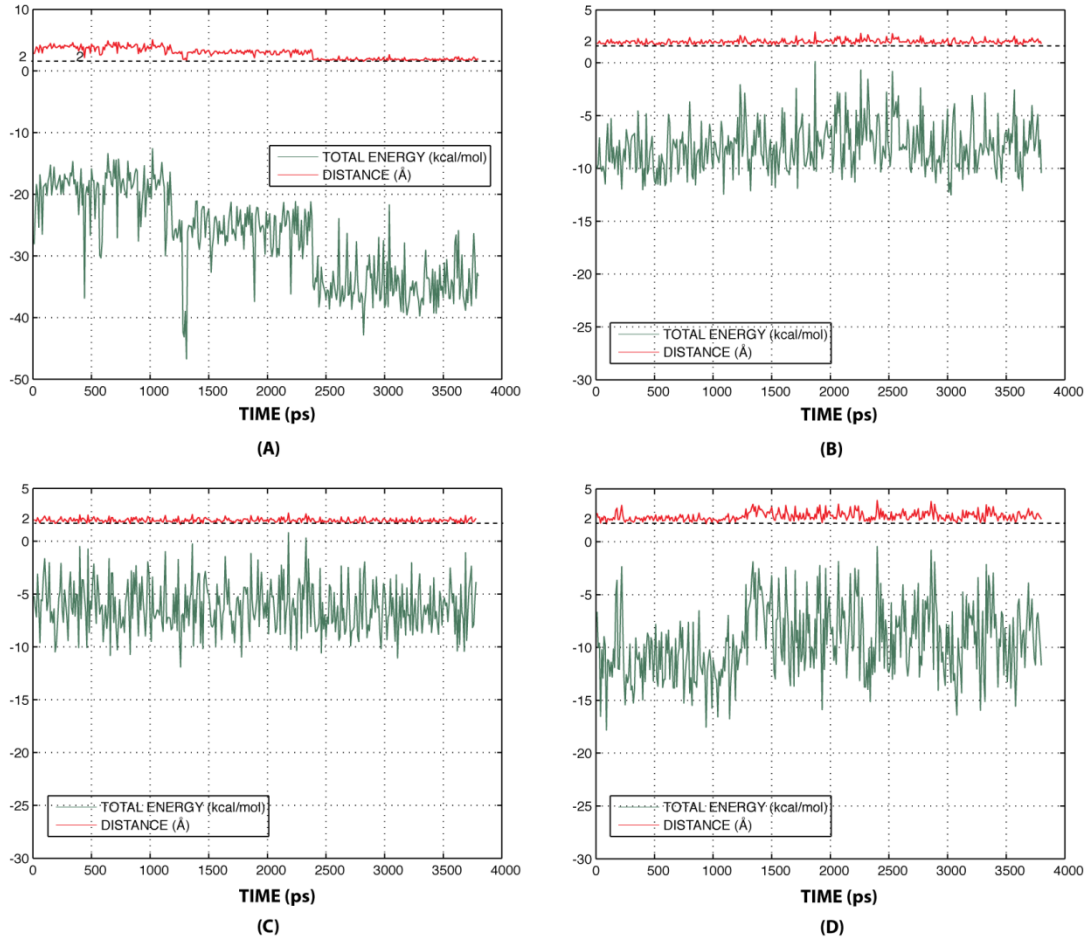


Figure 4.7 Time series graphs of the distance and interaction energy between (A) SER169 and the PO5 oxygen of the uracil nucleotide, (B) GLN144 and the O2 oxygen of uracil, (C) PHE158 and the O4 oxygen of uracil and (D) ASN204 and the uracil base.

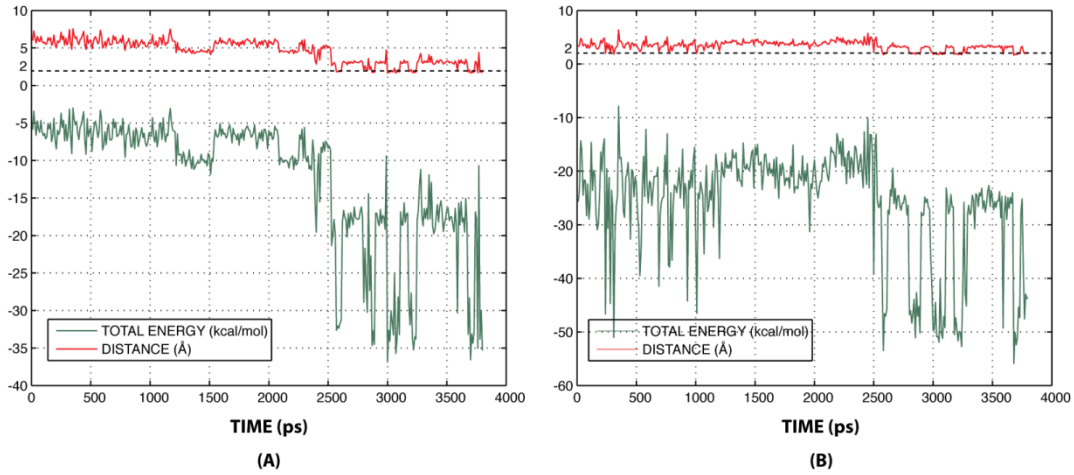


Figure 4.8 The time series of the distance and interaction energy between (A) HIS268 and WATER A and (B) WATER A and the O2 oxygen of the uracil nucleotide.

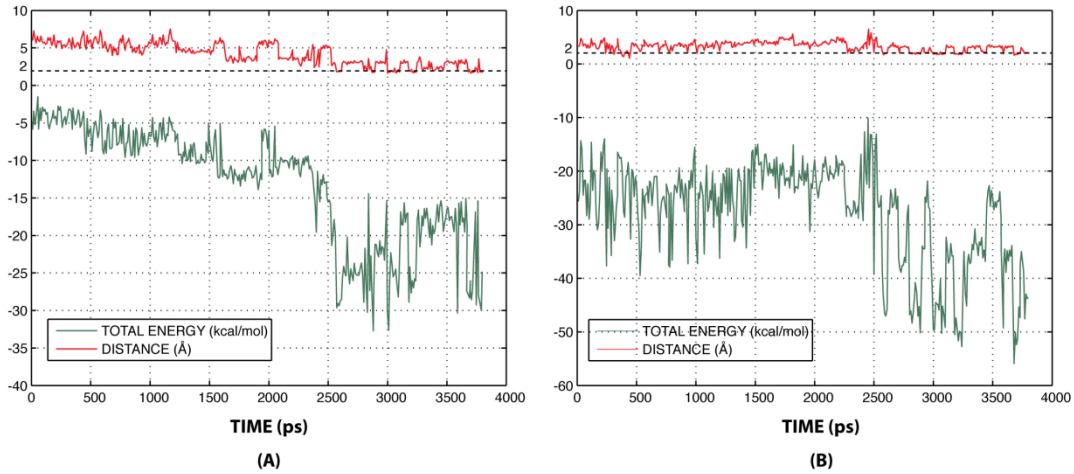


Figure 4.9 The time series of the distance and interaction energy between (A) HIS148 and WATER B and (B) WATER B and the PO3 oxygen of the uracil nucleotide.

There are quite a few amino acids that interact with the DNA double helix predominantly through electrostatic interaction (Figure 4.10). Serine (SER270) forms a water-mediated interaction with the oxygen of the phosphate group between adenine and the extrahelical uracil base. Serine (SER169) has an average interaction energy of -30.49 kcal.mol^{-1}, which is the strongest interaction with the DNA compared to the rest of the amino acids. SER169 forms a hydrogen bond with an oxygen between the extrahelical uracil base and a phosphate. Tyrosine (TYR248) forms a hydrogen bond with an oxygen on the phosphate backbone. Glutamine (GLN152), histidine (HIS212), proline (PRO167 and PRO168) and leucine (LEU272) forms nonspecific interaction with the DNA double helix. The combined effect of the amino acids stabilises the DNA in this unusual extrahelical conformation.

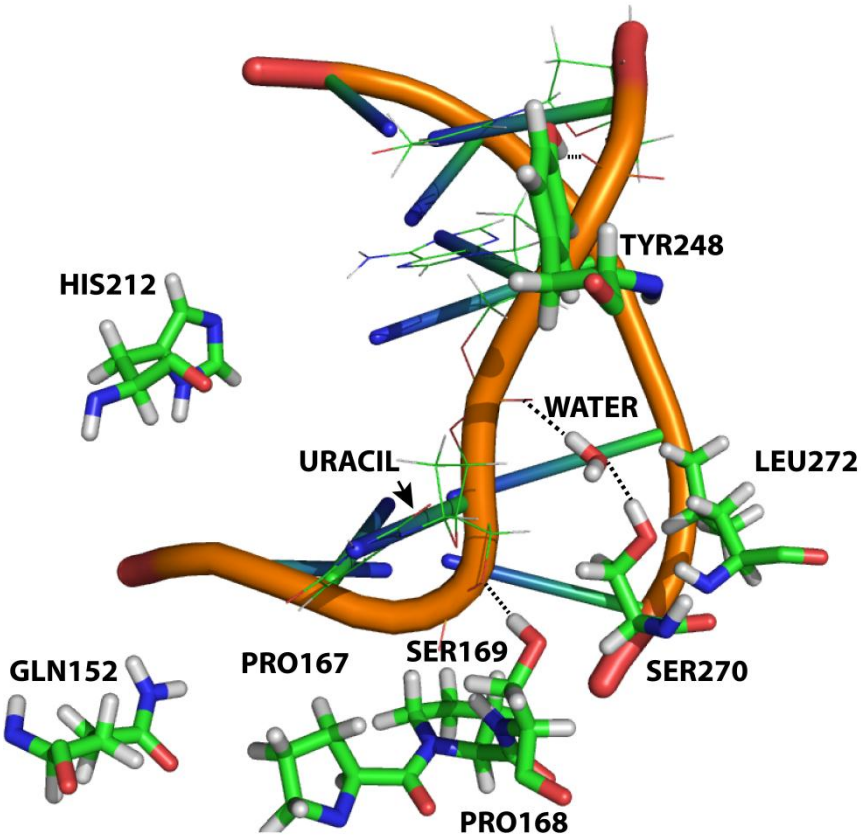


Figure 4.10 Amino acids of the hUDG enzyme interacting with the sugar-phosphate backbone of the DNA substrate.

4.4.3 Uracil nucleotide binding in hsvUDG

There are fewer interactions between the DNA substrate in hsvUDG protein than there are in the hUDG protein. Figure 4.11 illustrates the 2-dimensional interaction profile between the hsvUDG enzyme and the uracil nucleotide. This can be attributed to the fact that the DNA did not enter the binding pocket of the hsvUDG protein as deeply as it did in hUDG. The amino acids in Table 4.2 are arranged according to various regions of interaction with the DNA. The primary responsibility of the interactions between the amino acids and the DNA substrate is for the stabilisation of the substrate within the binding pocket. Amino acids in regions A, B and C are located near the extrahelical base, the deoxyribose sugar of the uracil and the rest of the DNA respectively.

Asparagine (ASN204) forms a water-mediated interaction with the oxygen of C4 on uracil. The carboxamide functional group of ASN204 forms two electrostatic interactions with the WATER A which in turn forms a hydrogen bond with the oxygen. The NH bond of the glycosidic bond between phenylalanine (PHE101) and alanine (ALA100) forms a hydrogen bond with the O4 oxygen on uracil just as the CYS157 and PHE158 sequence does in hUDG. Aspartate (ASP88) forms two interactions with the uracil nucleotide. The carboxyl functional group forms a hydrogen bond with the H3 hydrogen of uracil and a water-mediated interaction with the O2 oxygen of uracil. No significant interactions were made between the deoxyribose sugar and the protein.

Figures 4.12, 4.13 and 4.14 show the interaction energy time series of the water-mediated interaction of the amino acid residues ASN147, ASP88 and HI91 respectively, with the uracil nucleotide, as shown in Figure 4.11. From these figures, the correlation between the distance and the interaction energy of the water molecules from the amino acids and the uracil nucleotide can be seen. This drop and rise in interaction energy is caused by the location of the water molecule as it moves towards and away from the substrate. Figure 4.15 shows the time series graphs of the distance and interaction energy between the PHE101 and the O4 oxygen of uracil, and ASP88 and the H3 hydrogen of uracil.

Lysine (LYS216), serine (SER215) and leucine (LEU214) are responsible for latching on to the DNA double helix by inserting themselves into the groove of the helix (Figure 4.16). LYS216 has an average interaction energy of -80.8128 $kcal.mol^{-1}$. Electrostatic interactions accounts for the majority of this energy. No specific interaction between any single atoms can be identified. LYS216 is in close proximity to two ADENINE:THYMINE base pairs. LEU214 seems to insert itself into the empty space left by the extrahelical uracil base and producing an interaction energy of -17.76$kcalmol^{-1}$. This interaction in the hsvUDG protein seems to be responsible for preventing the uracil from flipping back into its original position and base pairing with the guanine base.

REGION	AA	Average Interaction Energy (kcal.mol^{-1})											
		Uracil Nucleotide				Uracil Base				Complete DNA Substrate			
		DISTANCE	VDW	ELEC	TE	DISTANCE	VDW	ELEC	TE	DISTANCE	VDW	ELEC	TE
A	PRO89	8.65	-1.46	-2.71	-4.17	6.68	-1.31	-0.05	-1.36	20.31	-1.47	-3.06	-4.53
	TYR90	6.38	-3.92	-4.50	-8.42	4.58	-3.46	-1.59	-5.05	17.49	-3.99	-6.37	-10.35
	HIS210	6.65	-3.05	-2.24	-5.29	6.61	-1.89	-0.13	-1.76	12.25	-6.74	-13.50	-20.24
	ASN147	10.81	-0.21	-4.21	-4.42	8.23	-0.34	-5.34	-5.68	21.34	-0.54	-6.23	-6.77
	PHE101	8.31	-2.34	-1.80	-4.14	5.42	-2.23	-1.10	-3.33	17.94	-2.38	-2.54	-4.92
	ASP88	5.80	-2.50	-6.63	-9.13	5.77	-0.89	-2.68	-3.57	13.54	-4.67	-2.70	-7.37
	SER112	6.14	-1.32	-5.00	-6.32	5.96	0.79	-1.29	-2.08	12.25	-1.65	-10.44	-12.09
B	TOTAL A	-	**-14.80**	**-27.09**	**-41.89**	-	**-10.65**	**-12.18**	**-22.83**	-	**-21.43**	**-44.84**	**-66,27**
	GLN87	7.83	-0.82	-5.64	-6.46	6.93	-0.55	0.019	-0.53	16.38	-1.83	-8.27	-10.10
	HIS91	7.02	-1.71	-3.85	-5.56	7.86	-0.30	-0.62	-0.92	16.44	-1.99	-1.13	-3.12
C	TOTAL B	-	**-2.52**	**-9.49**	**12.02**	-	**-0.86**	**-0.60**	**-1.46**	-	**-3.82**	**-9.40**	**-13.22**
	ALA157	8.31	-0.40	-3.25	-3.65	10.19	-0.04	-0.47	-0.51	14.98	-0.52	-5.20	-5.72
	SER212	7.00	-0.87	-3.51	-4.38	8.43	-0.15	0.05	-0.1	7.90	-3.25	-6.80	-10.05
	PRO213	9.97	-0.16	-0.50	-0.66	11.47	0	0	0	7.66	-3.78	0.12	-3.66
	LEU214	9.75	-0.39	-2.32	-2.71	12.39	0	0	0	2.53	-14.17	-3.59	-17.76
	SER215	10.57	-0.05	-1.41	-1.46	12.30	0	0	0	6.13	-4.79	-7.31	-12.10
	LYS216	15.59	0	-0.06	-0.06	17.61	0	0	0	7.13	-7.51	-74.05	-80.81
	PRO111	8.62	-0.37	-1.55	-1.92	9.40	-0.08	-0.17	-0.25	11.85	-2.15	-6.97	-9.12
	PRO211	7.87	-0.75	-3.10	-3.85	7.31	-0.37	-1.10	-1.47	12.71	-1.69	-7.02	-8.71
	SER209	12.21	-0.02	-0.93	-0.95	12.39	-0.01	-0.08	-0.09	13.18	-1.94	-14.90	-16.84
	VAL217	14.65	-0.00	-0.06	-0.06	15.61	0	0	0	10.85	-1.21	0.69	0.52
	TOTAL C	-	**-3.00**	**-10.48**	**-14.97**	-	**-0.66**	**-1.28**	**-2.19**	-	**-37.62**	**-110.99**	**-164.26**
TOTAL PROTEIN		**14.01**	**-20.97**	**-29.22**	**-50.19**	**11.01**	**-13.58**	**-7.74**	**-21.32**	**24.48**	**-68.48**	**-182.72**	**-251.20**

Table 4.2 Average interaction energies between the substrate and key amino acid residues in the hsvUDG enzyme. Three regions of the substrate are considered. These regions include the uracil base, the uracil nucleotide and the complete DNA substrate. Amino acids in region A, B and C are located near the extrahelical uracil base, the deoxyribose sugar of the uracil nucleotide and the rest of the DNA respectively.

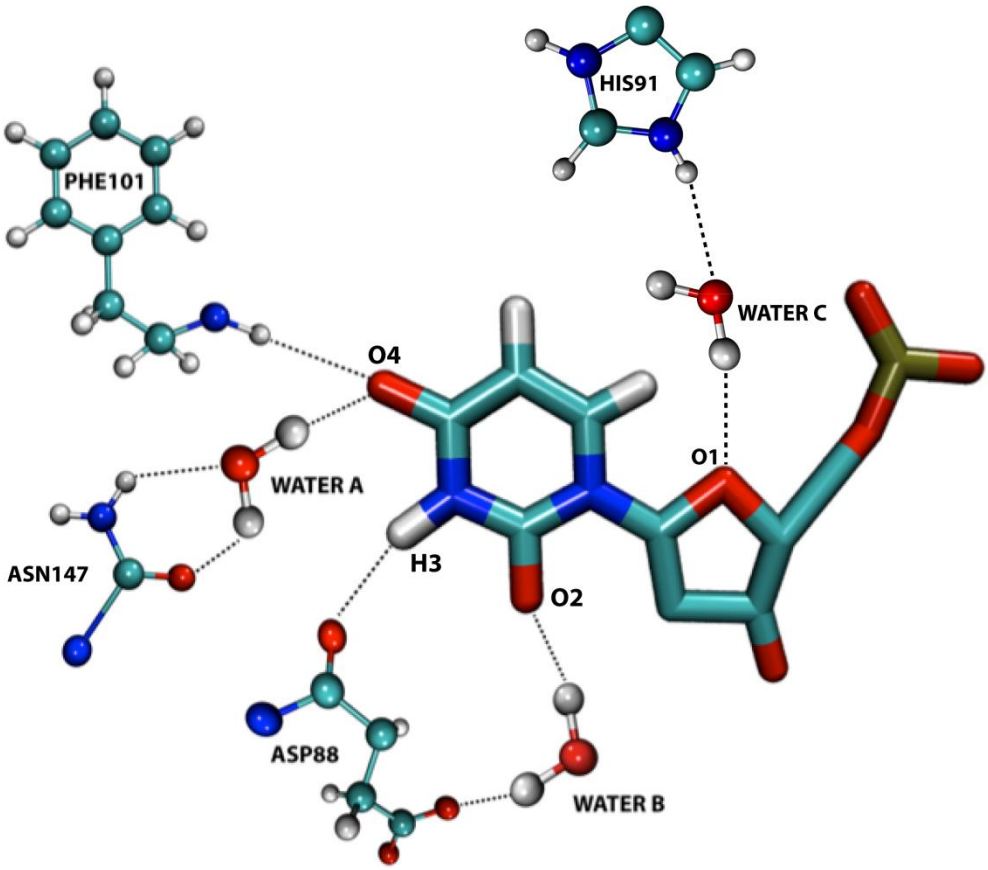


Figure 4.11 2-dimensional illustration of significant interactions between the amino acids in the binding pocket of the hsvUDG enzyme and the uracil nucleotide.

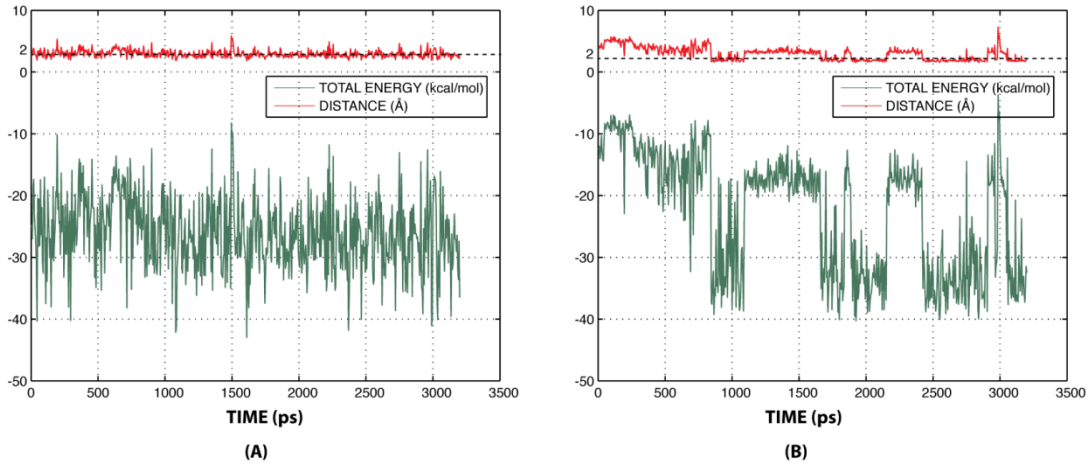


Figure 4.12 The time series of the distance and interaction energy between (A) ASN147 and WATER A, and (B) WATER A and the O4 oxygen of the uracil nucleotide.

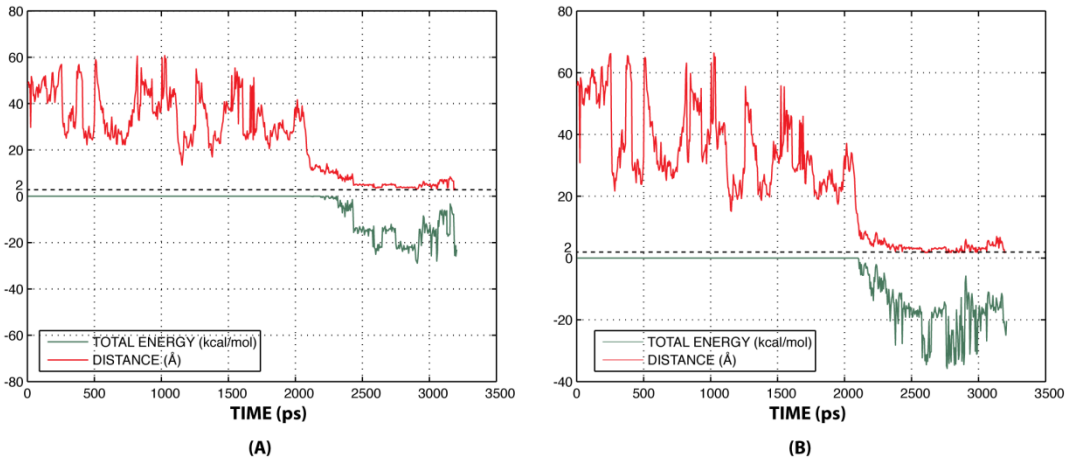


Figure 4.13 The time series of the distance and interaction energy between (A) ASP88 and WATER B, and (B) WATER B and the O2 oxygen of the uracil nucleotide.

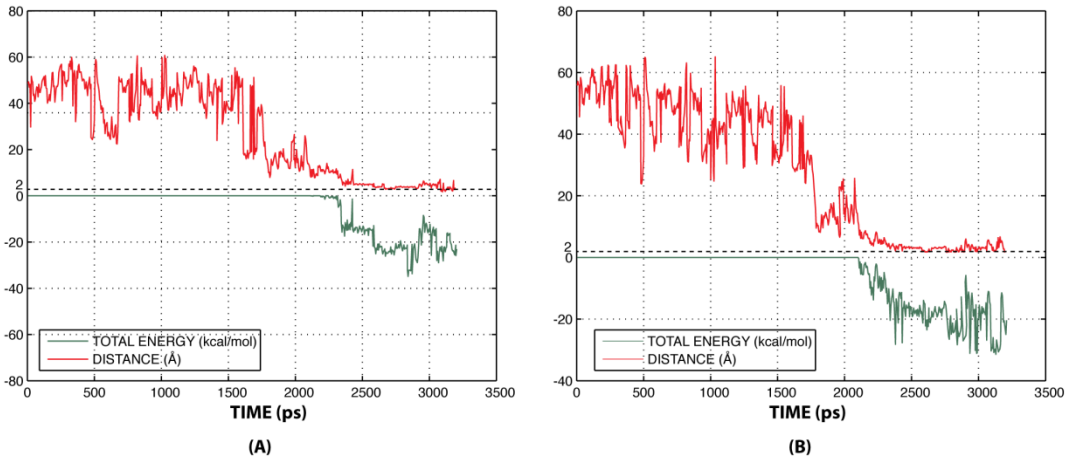


Figure 4.14 The time series of the distance and interaction energy between (A) HIS91 and WATER C, and (B) WATER C and the O1 oxygen of the uracil nucleotide.

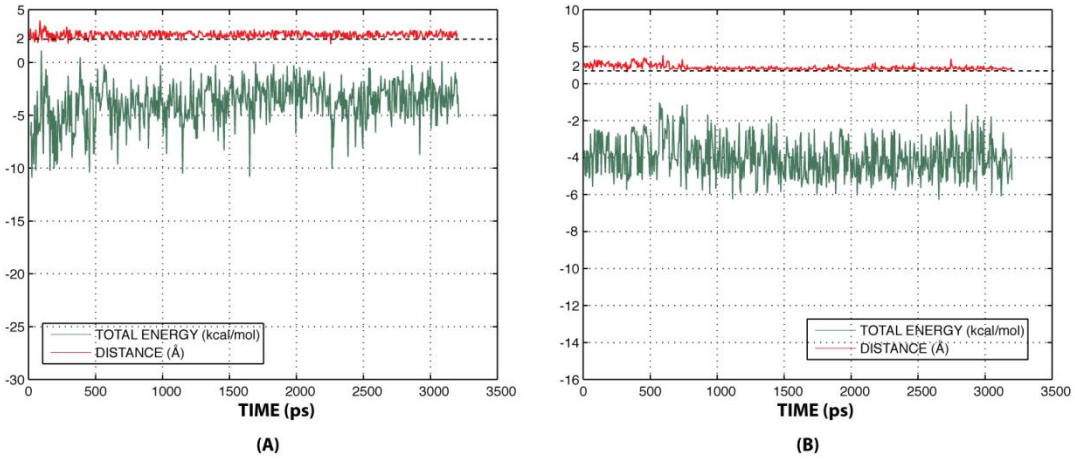


Figure 4.15 Time series graphs of the distance and interaction energy between (A) PHE101 and the O4 oxygen of uracil, and (B) ASP88 and the H3 hydrogen of uracil.

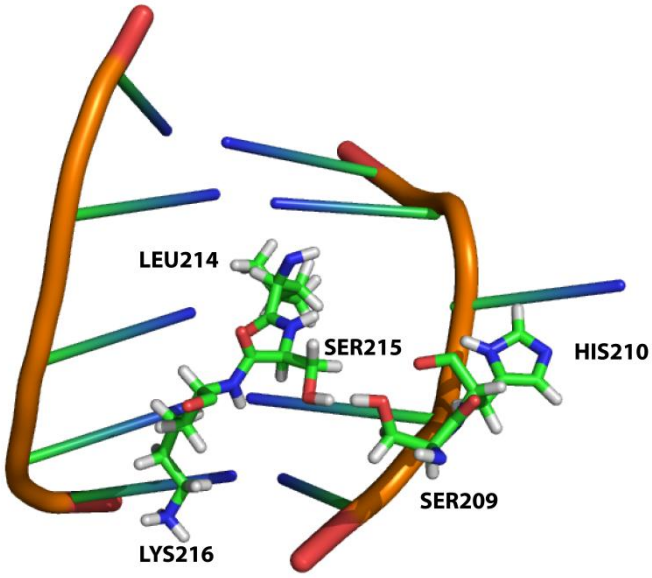


Figure 4.16 Amino acids of the hsvUDG enzyme interacting with the sugar phosphate backbone of the DNA substrate.

4.5 Comparing the Behaviour of the hUDG and hsvUDG Enzymes

The hUDG protein has approximately 26kcal.mol^{-1} greater interaction energy with the DNA substrate than the hsvUDG protein. There are several reasons that were identified as to why this is the case. Firstly, the geometric dimensions of the secondary structure of the proteins need to be considered. All geometric data used are measured averages over the 7ns production simulation.

From Figures 4.17 (B) and 4.18 (B), it can be seen that hsvUDG is slightly wider along the z-axis, than hUDG. The angle II displayed by the figures, is the angle between the β sheet (green) and the α helix (pink) indicated by the blue arrows in Figure 4.17 (B) and Figure 4.18 (B). The angle II in hsvUDG is greater than the same angle in hUDG by 9°. This difference in separation is in agreement with the difference in width between the two proteins. The width along the x-axis of the hUDG protein is more than that of the hsvUDG protein (Figure 4.17 (A) and Figure 4.18 (A)). Although these proteins are of the same evolutionary family, namely the uracil-DNA glycosylase superfamily, there are slight differences in the amino acid complement between them. These differences in amino acids bring about differences in the interactions within the proteins. This is clearly visible from the difference in the geometric data calculated.

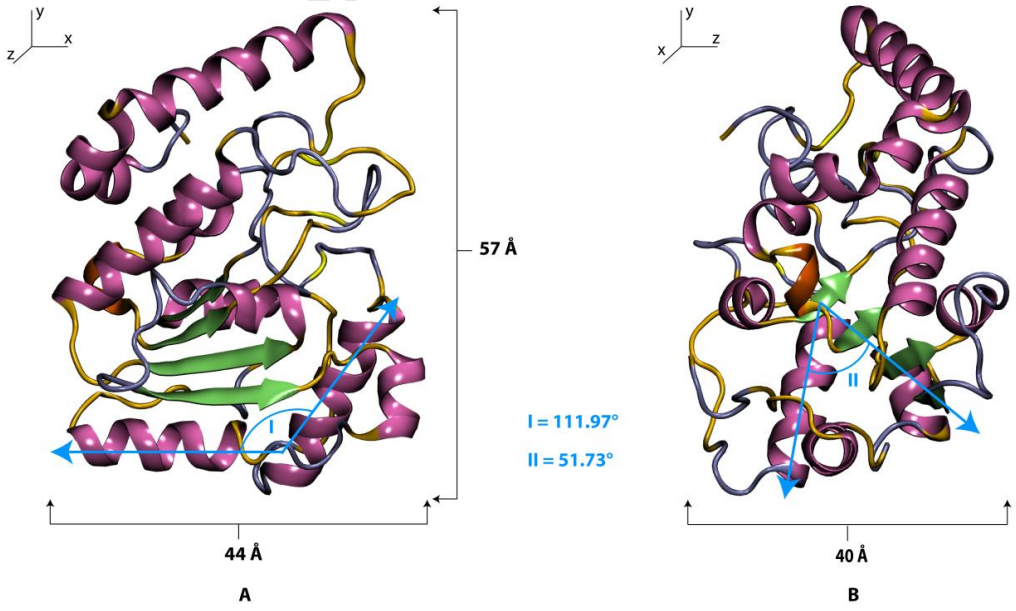


Figure 4.17 Geometric information for the hUDG protein. (A) Front and (B) side view.

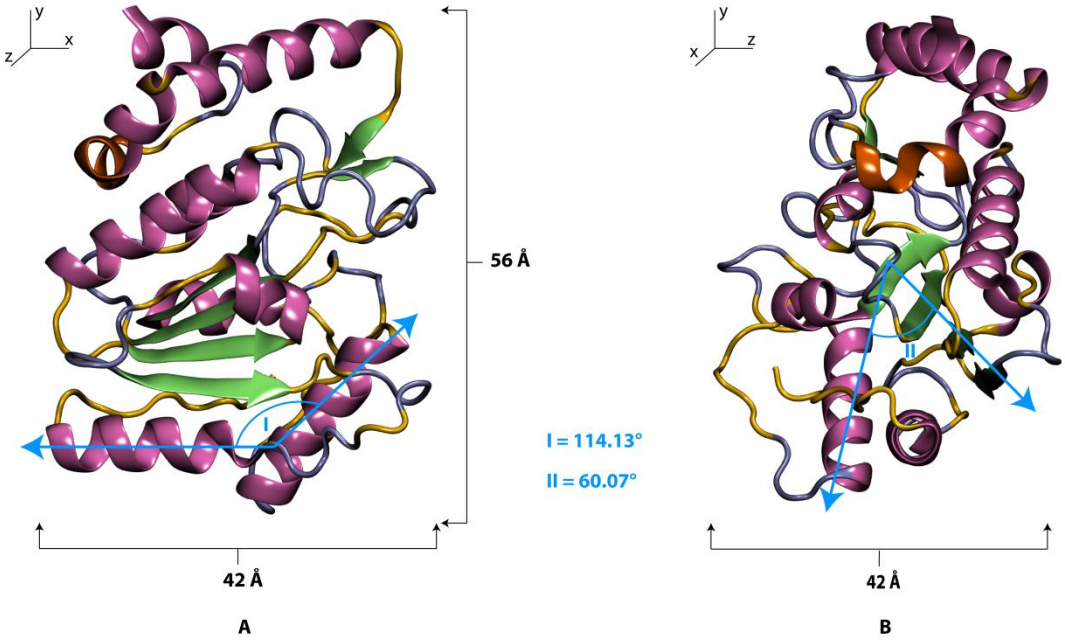


Figure 4.18 Geometric information for the hsvUDG protein. (A) Front and (B) side view.

Further differences in the electrostatics of the proteins can be seen. hUDG has a more positive binding pocket than that of hsvUDG (Figure 4.19). Field lines are included in Figure 4.19 in order to illustrate the difference in the field strength created by the arrangement of the amino acids. From Figure 4.19, the arrangement of positively charged amino acids at the binding pockets can be seen, and more neutral to negatively charged amino acids can be seen around the rest of the enzymes. A positive binding pocket is essential for these DNA binding enzymes. The highly negatively charged sugar-phosphate backbones of the DNA provide an ideal electrostatic interaction potential for positively charged binding pockets. From Table 4.1 we can see that there are several amino acids that have large electrostatic interactions with the DNA in the hUDG enzyme. SER169 from hUDG has an interaction energy of -30.4875kcal.mol^{-1} of which 29.9643kcal.mol^{-1} is due to electrostatic interactions. LYS216 from hsvUDG has an interaction energy of -80.8128kcal.mol^{-1} of which -74.0481kcal.mol^{-1} is due to electrostatic interactions (Table 4.2). These are just two examples of the role electrostatic interactions play in stabilising the DNA substrate in the binding pocket. LEU272 and LEU214 are conserved amino acids in hUDG and hsvUDG respectively,

however only LEU214 inserts itself into the space left by the extrahelical uracil base. This difference in behaviour can be explained by the difference in geometric behaviour and deeper penetration of the DNA substrate in the hsvUDG enzyme caused by more favourable interactions.

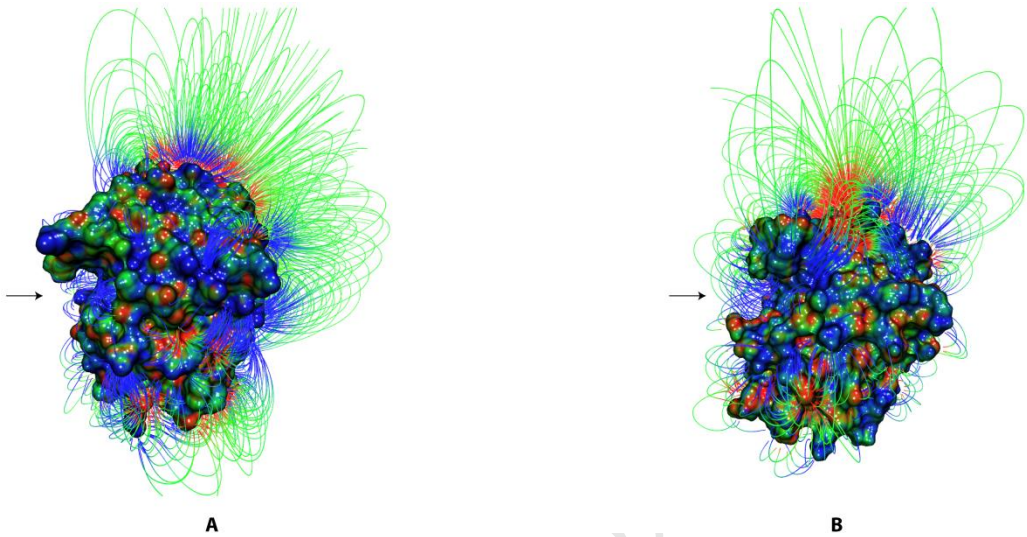


Figure 4.19 Electrostatic surface potential and field lines of (A) hsvUDG and (B) hUDG. Arrows indicate the binding pockets of the enzymes. Blue, green and red indicate the positive, neutral and negative areas of the surfaces and field lines.

Due to the above mentioned differences between hsvUDG and hUDG, very different initial DNA binding behaviour can be seen. Figure 4.20 shows the secondary structure conformation of the proteins around the DNA substrate. The DNA in the figures were aligned with each other and shown in the same angle in Figure 4.20 (A) and (B). This was done so that it can be clearly seen how differently the proteins wrap around the DNA. hUDG covers more of the DNA substrate and the uracil enters the binding pocket much deeper in hUDG than in hsvUDG.

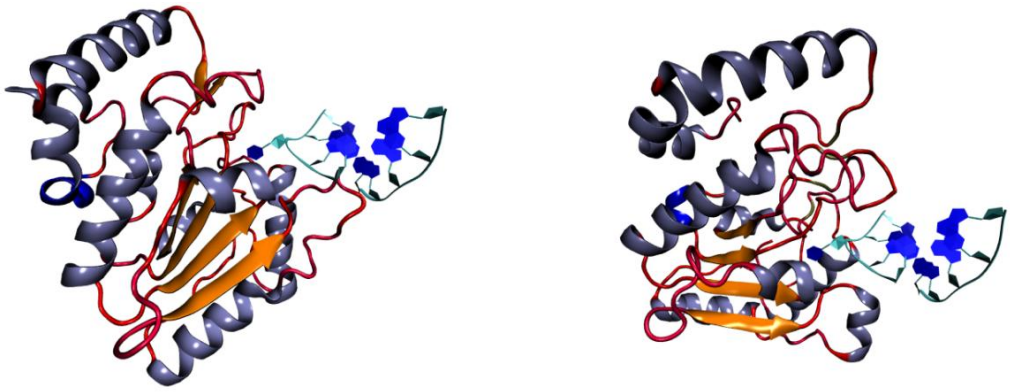


Figure 4.20 The overall manner by which (A) hsvUDG and (B) hUDG wrap around the DNA substrate.

4.6 Discussion

From the data presented it can be seen that although the two proteins investigated (hsvUDG and hUDG) are evolutionary quite similar in their amino acid complement, their behaviour towards a extrahelical uracil-DNA substrate shows how differently they carry out their function. hUDG seems to interact more favourably with the DNA substrate than hsvUDG. Based on the distance between the uracil head and the ASN204 and ASN147 amino acids (which are considered the base of the binding pocket), it can be seen that the uracil in hUDG enters the binding pocket deeper than in hsvUDG (Table 4.3). The interaction energy between the substrate and the protein is greater in hUDG. hsvUDG is approximately 22Å bigger than the binding pocket volume than hUDG.

Summary of Data for DNA-Protein Simulations								
	Average Volume of Binding pocket ($Å^3$)	DNA Interaction Energy ($Kcal.mol^{-1}$)	Angle I (°)	Angle II (°)	Front Face Width (Å)	Side Face Width (Å)	Height (Å)	Uracil Distance From Binding-Pocket/(ASN204|147) (Å)
hsvUDG	291	-251.4696	114.13	60.07	42	42	57	5.07
hUDG	269	-278.3699	111.97	51.73	44	40	56	2.14

Table 4.3 Summary of all the data compiled for the hDUG and hsvUDG proteins.

These differences are due to the different amino acid complement in each protein, which brings about different interactions between amino acids, which ultimately affects the geometric shape of the protein as a whole.

References:

1. T. Lindahl, *Nature*, 1993, **362**, 709-715.
2. L. H. Pearl, *Mutation Research-DNA Repair*, 2000, **460**, 165-181.
3. S. R. W. Bellamy, K. Krusong and G. S. Baldwin, *Nucleic Acids Res.*, 2007, **35**, 1478-1487.
4. C. Cao, Y. L. Jiang, K. D. J. and J. T. Stivers, *J Am Chem Soc.*, 2006, **128**, 13034-13035.
5. C. Cao, Y. L. Jiang, J. T. Stivers and F. Song, *Nat. Struct. Mol. Biol.*, 2004, **11**, 1230-1236.
6. I. Wong, A. J. Lundquist, A. S. Bernards and D. W. Mosbaugh, *J. Biol. Chem.*, 2002, **277**, 19424-19432.
7. K. Krusong, E. P. Carpenter, B. S. Watson, R. Savva and G. S. Baldwin, *Journal of Biological Chemistry*, 2006, **281**, 4983-4992.
8. A. J. Drummond, B. Ashton, S. Buxton, M. Cheung, A. Cooper, J. Heled, M. Kearse, R. Moir, S. Stonees-Havas, S. Sturrock, T. Thierer and A. Wilson, 5th edn., 2010.
9. B. R. Brooks, R. E. Bruccoleri, D. J. States, S. Swaminathan and M. Karplus, *Journal of Computational Chemistry*, 1987, **4**, 187-217.
10. N. Foloppe and A. D. MacKerell, *Journal of Computational Chemistry*, 2000, **21**, 86-104.
11. A. D. MacKerell and N. K. Banavali, *Journal of Computational Chemistry*, 2000, **21**, 105-120.
12. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *Journal of Chemical Physics*, 1983, **79**, 926-935.
13. A. R. Leach, *Molecular Modelling: Principles and Applications*, Prentice Hall, New York, 2001.
14. P. P. Ewald, *Annals of Physics*, 1921, **64**, 253-287.
15. A. Y. Toukmaji and B. J. A., *Computer Physics Communications*, 1996, **95**, 73-92.
16. S. S. Parikh, C. D. Mol, G. Slupphaug, S. Bharati, H. E. Krokan and J. A. Tainer, *Embo Journal*, 1998, **17**, 5214-5226.
17. R. Savva, K. McAuleyhecht, T. Brown and L. Pearl, *Nature*, 1995, **373**, 487-493.
18. J. E. Nielsen, *J. Mol. Graph. Model*, 2006, **25**, 691-699.
19. W. P. Jencks, *Catalysis in Chemistry and Enzymology*, New York, 1975.
20. A. Fersht, Freeman, New York, 4th edn., 1985.
21. K. Morokuma, *Accounts of Chemical Research*, 1977, **10**, 294-300.
22. N. M. Luscombe, R. A. Laskowski and J. M. Thornton, *Nucleic Acids Research*, 2001, **29**, 2860-2874.
23. M. P. Waller, A. Robertazzi, J. A. Platts, D. E. Hibbs and P. A. Williams, *Journal of Computational Chemistry*, 2006, **27**, 491-504.
24. B. Kavli, G. Slupphaug, C. D. Mol, A. S. Arvai, S. B. Peterson, J. A. Tainer and H. E. Krokan, *Embo Journal*, 1996, **15**, 3442-3447.
25. I. Y. Torshin, I. T. Weber and R. W. Harrison, *Protein Engineering*, 2002, **15**, 359-363.

Chapter 5

Identifying differences in inhibitor interactions between hUDG and hsvUDG

5.1 Introduction

Uracil DNA glycosylase is known to be responsible for the removal of uracil from DNA by the cleavage of the *N*-glycosidic bond.[1] However, the functional role of the protein in certain viruses, such as the herpes simplex virus, has yet to be resolved. It has been hypothesised that hsvUDG plays a role in the reactivation of the herpes simplex virus (HSV)[2] and for efficient replication in nerve tissue.[3] The lack of cellular UDG in neurons combined with the continual deamination of cytosine creates an environment where the need for viral UDG is a necessity for proliferation of the virus.[4] To gain an understanding of the functional role of the hsvUDG enzyme in the virus, the activity of the enzyme has to be suppressed. Inhibitors are ideal for the suppression or complete inactivation of enzymes. Several inhibitors have been developed for hsvUDG. Given that critical amino acids are conserved between hsvUDG and hUDG, (Figure 5.1) and the overall secondary structure of the proteins are extremely similar, very few inhibitors have been able to selectively inhibit hsvUDG effectively.[5]

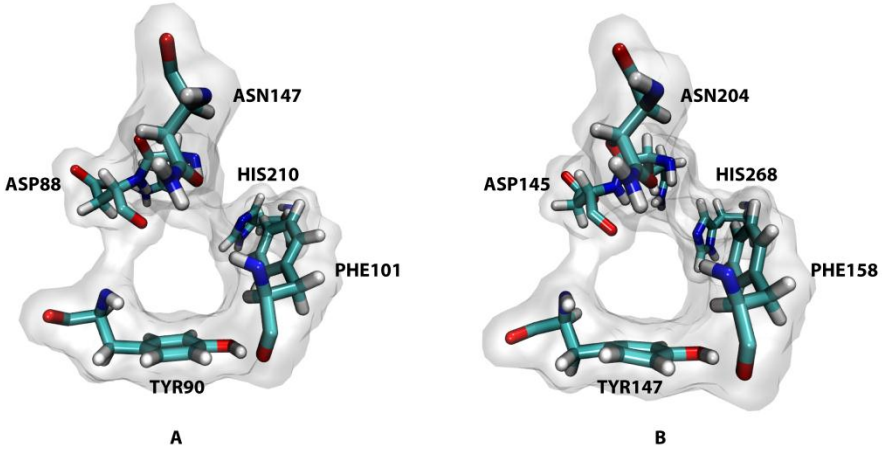


Figure 5.1 Conserved amino acids in binding pocket of (A) hsvUDG and (B) hUDG.

5.2 General Structure of Inhibitors

The 6-(4-alkylanilino)-uracil inhibitors are synthesised analogues of the uracil substrate. These compounds were found to be competitive with DNA as inhibitors of hsvUDG. The structures of the inhibitors investigated in this thesis are shown in Figure 5.2. The IC_{50} values can be seen in Table 5.1. The binding model of the inhibitors shown in Figure 5.3 was proposed due to the strong hydrophobic character of the alkyl chain and its interaction with the "hydrophobic cleft" created by the amino acids proline (PRO111 and PRO213) and leucine (LEU214).[6]

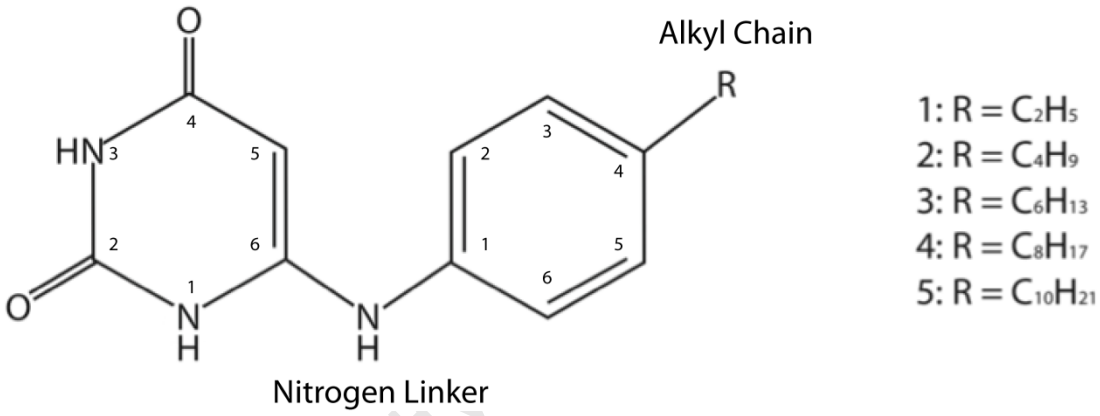


Figure 5.2 The molecular structure of the 6-(4-alkylanilino)-uracil inhibitors investigated.

Inhibitor	IC_{50} (μM)
1	500
2	150
3	30
4	8
5	35

Table 5.1 Inhibitors and their respective IC_{50} values.

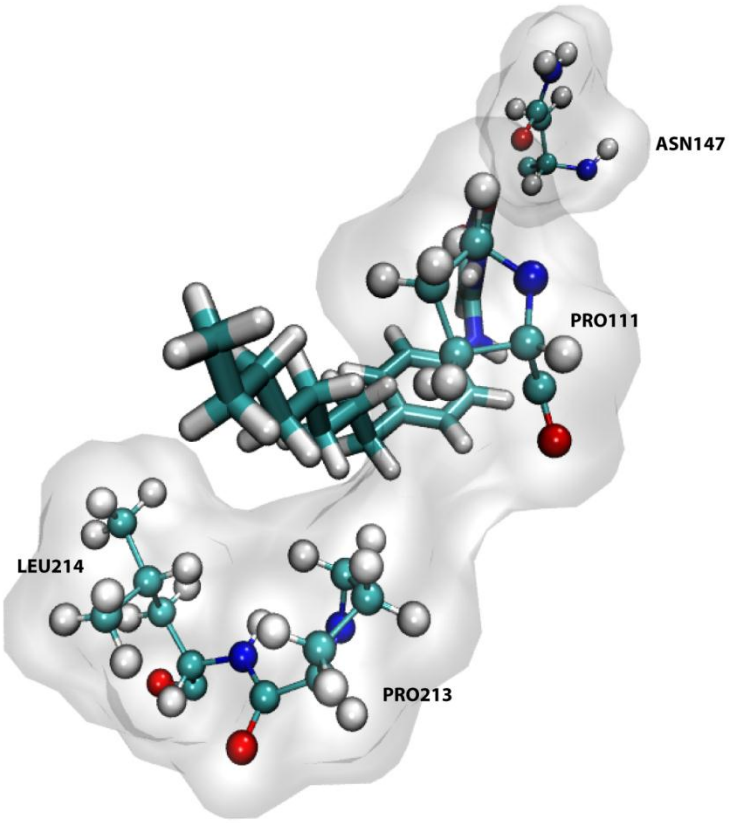


Figure 5.3 Proposed binding model of 6-(4-octylanilino)-uracil.[6]

5.3 Parameterisation of Inhibitors

As indicated in chapter 2, molecular mechanics is expressed through the force field description of a molecule. Force fields involve bonded and nonbonded terms for all atoms that comprise the molecule. These force field parameters are then used by the potential energy function to produce the mechanics of the molecule of interest. Before simulations can be performed, it is necessary to ensure that all the parameters for the types of bonds, angles, torsion angles, improper angles and nonbonded interactions exist and work with the system of interest. For the 6-(4-alkylanilino)-uracil inhibitors, the charges for NN2U and HN2 and the dihedral parameters Φ and Ψ (Figure 5.4) were not present and were parameterised using the methods described below.[7]

All the empirical calculations were carried out using the CHARMM program[8] using a dielectric constant of 1.0. The CHARMM-modified TIP3P[9] water model

was used in all calculations. QM calculations were carried out with the GAUSSIAN 03 program.[10]

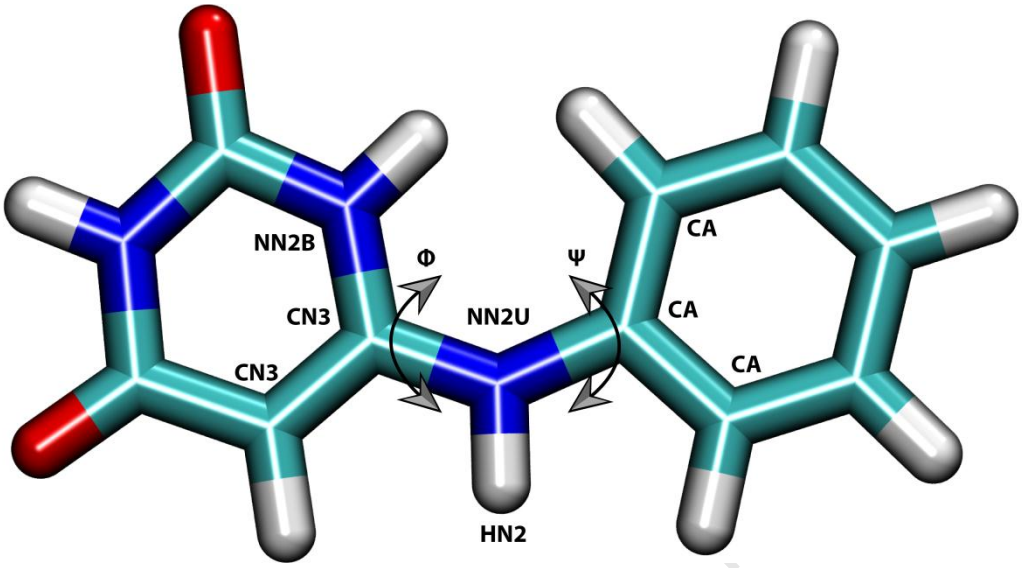


Figure 5.4 Definition of torsion angles characterising the nitrogen linkage between the uracil and benzene ring structures with atom names.

5.3.1 Charge Parameterisation

In order to stay consistent with the general parameterisation of charge, the procedure carried out by MacKerell et al[11] was followed. Minimum interaction energies and geometries between model compound and water were determined by optimising the intermolecular distance at the HF/6-31G(d) level of theory while constraining the model compound at the HF/6-31G(d) optimised geometry and the water at the TIP3P internal geometry.[9] The orientation of the water molecule as the distance was varied can be seen in Figure 5.5. The partial charges of NN2U and HN2 were adjusted in the empirical force field in order to reproduce minimum interaction energies and distances obtained quantum mechanically, as closely as possible, using the CHARMM27 force field. Interaction orientations were identical to those used in the QM calculations (Table 5.2). Model compound-water interaction energies were scaled by a factor of 1.16 and distances were offset by -0.2 Å. The QM interaction energy was determined as the total energy of the supermolecular complex minus the sum of the monomer energies.[11]

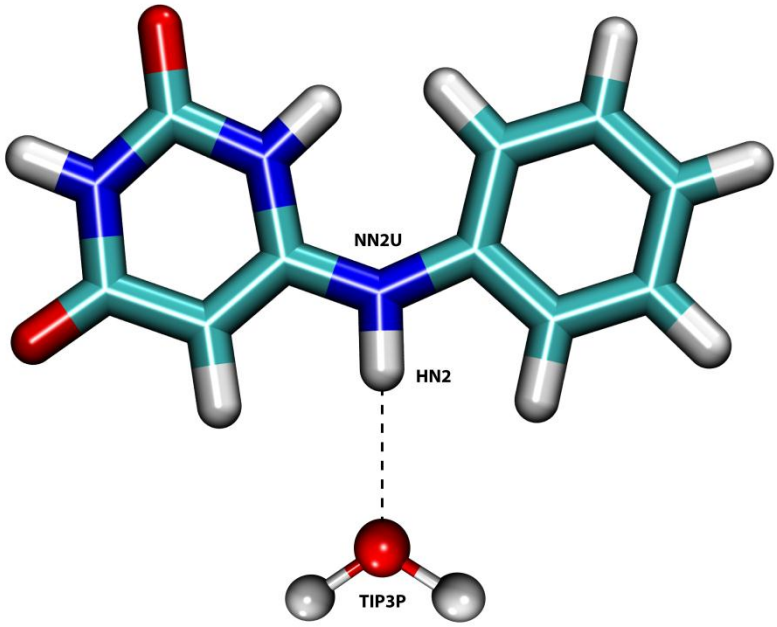


Figure 5.5 Interaction orientation of the model molecule and the TIP3P water molecule.

	Interaction Energies (kcal/mol)		Interaction Distances (Å)	
	Empirical	QM	Empirical	QM
HN2--OH_2	-3.60	-3.72	2.01	2.14

Table 5.2 Minimum water interaction energies and distance.

5.3.2 Dihedral Parameterisation

Figure 5.4 shows the Φ and Ψ dihedral angles that were parameterised. Similar to the charge parameterisation method, the parameters of the dihedral in the empirical force field were modelled to reproduce rotational plots of the same molecule using QM. Initial parameters were obtained from molecules that had similar configurations. No truncation of nonbonded interaction was used in the empirical calculations.

The above method was achieved by varying the dihedral angle concerned through 360° at 10° intervals, constraining the selected dihedral with a force constant of 10,000kcal/mol/degree^2, minimising using 200 steps of steepest decent followed by 200 steps of Newton-Raphson minimisation methods and

measuring the energy of the molecule.[12] The energies obtained were plotted against their respective dihedral angles to create a molecular mechanical (MM) rotational plot. Dihedral energy surfaces were produced and geometries were optimised at the HF/6-31G(d) level of theory and were conducted for each of the resulting MM dihedral points. The empirical force field parameters (Table 5.3) were adjusted in order for the MM rotational plot to be fitted to the QM rotational plot (Figure 5.5.1).

ANGLE TYPE	K_ϕ	n	γ
NN2B-CN3-NN2U-CA	1.6	2	180
CN3-NN2U-CA-CA	1.5	2	180

Table 5.3 Dihedral parameters for the nitrogen linker.

After an iterative parameterisation method going back and forth between charge and dihedral parameterisation, the final charge for specific atom names calculated are shown in Table 5.3. All van der Waals parameters were used from existing atom types that showed similar characteristics to the atoms in the molecules being parameterised.

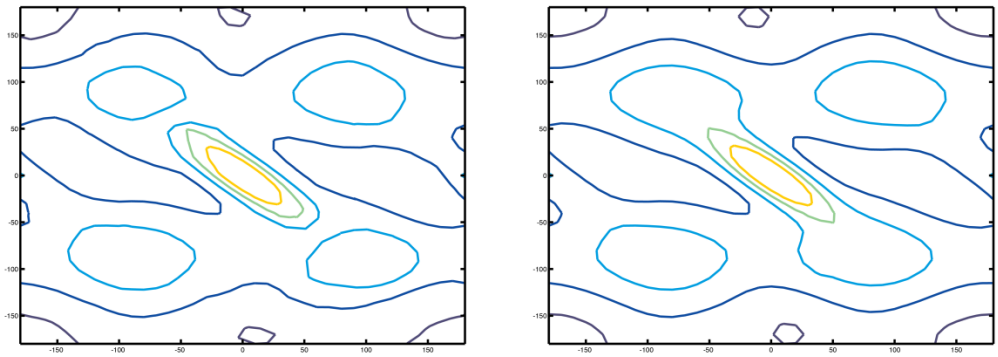


Figure 5.5.1 Rotational plots obtained using (A) the force field (MM) and (B) quantum mechanics (QM).

General Inhibitor		
Atom Name	Atom Type	Atomic Charge
C2	CN1T	0.55
O2	ON1	-0.45
N3	NN2U	-0.46
H3	HN2	0.36
C4	CN1	0.53
O4	ON1	0.48
C5	CN3	-0.15
H5	HN3	0.10
CG	CA	-0.115
HG	HP	0.115
CD1	CA	0.115
HD1	HP	0.115
CE2	CA	-0.115
HE2	HP	0.115
CZ	CA	-0.115
HZ	HP	0.115
N1	NN2B	-0.34
H1N	HN2	0.48
C6	CN3	0.20
N2	NN2U	0.59
H2N	HN2	0.365
CD2	CA	-0.115
CE1	CA	0.000
CA1	CTL2	-0.180
H1'	HAL	0.090
H1''	HAL	0.090
CA2	CTL2	-0.180
H2'	HAL	0.090
H2''	HAL	0.090
CA3	CTL2	-0.180
H3'	HAL	0.090
H3''	HAL	0.090
CA4	CTL3	-0.270
H4'	HAL3	0.090
H4''	HAL3	0.090
H4	HAL3	0.090

Table 5.4 Atom type and name and their partial atomic charges for the newly parameterised inhibitors.

5.4 Molecular Dynamics Simulations

Simulations were carried out on both hUDG and hsvUDG with all 5 inhibitors (Figure 5.2). The simulations were carried out using the CHARMM33b2[8] program that applies the empirical energy function mentioned in chapter 2. The proteins were modelled using the CHARMM27[11, 13] all-atom force field which was designed to simulate proteins and nucleic acids. The newly parameterised charge and dihedral parameters were used for the inhibitors. A 40Å radius TIP3P[9] water sphere which consisted of a 5Å buffer region and a 35Å radius dynamic region was used. Leapfrog langevin dynamics were used in all simulations. Water molecules with heavy atom distances within 3Å of the solute were removed. The switching function was used to account for the nonbonded interactions. The switching function was initiated at a cutoff distance of 10Å and truncated at 12Å from the atom concerned. All hydrogen bond lengths were kept constant using the SHAKE[14] algorithm. The nonbonded interaction list and solvent image were updated every 10fs and a group-by-group selection criteria was used for the inclusion lists. The water sphere was centred on the binding pocket of the protein to ensure that the nonbond cut-off values did not include any region beyond the water spheres boundary. The systems were first heated gradually from 145K to 300K and then equilibrated for 8ns at a pressure of 1bar and a temperature of 300K. This was followed by a 7ns production simulation. Data for both simulations were stored at 10ps intervals. The standard deviation in temperature fluctuation during the production period is ±1.4K.[15]

5.5 Free Energy Perturbation Procedure

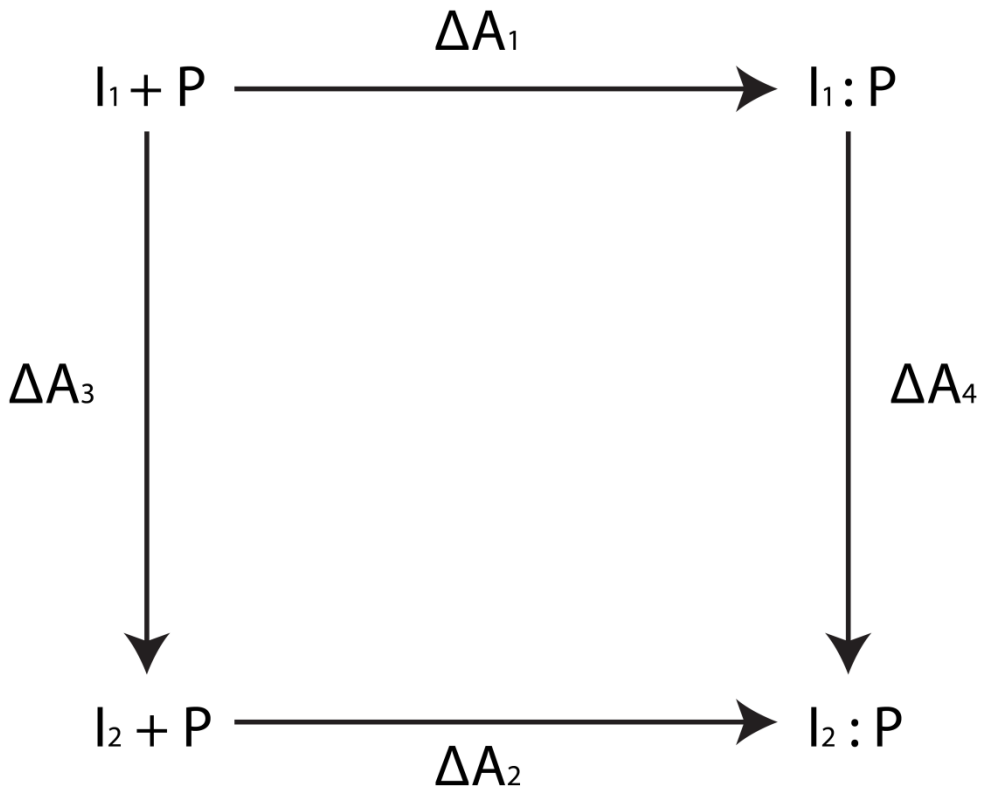


Figure 5.6 Thermodynamic cycle used to calculate ΔΔG.

This thermodynamic cycle is covered in chapter 3. Consider two ligands, I_1 and I_2, which could be inhibitors of an enzyme P. If ΔA_1 and ΔA_2 are the free energy of binding to the enzyme for inhibitors I_1 and I_2 respectively, then the relative binding affinity is $\Delta\Delta A = \Delta A_2 - \Delta A_1$. In order to simplify this calculation, we can consider using a thermodynamic cycle as shown in Figure 5.6. Because free energy is a state function, from Figure 5.6, $\Delta A_1 + \Delta A_4 = \Delta A_3 + \Delta A_2$. ΔA_3 corresponds to the free energy difference of the two ligands, I_1 and I_2 in solution, and ΔA_4 is the free energy difference of the two ligands, I_1 and I_2 in intermolecular (protein) complexes in solution. By rearranging the equation, $\Delta A_4 - \Delta A_3 = \Delta A_2 - \Delta A_1$. Therefore, computing $\Delta A_4 - \Delta A_3$ allows for the evaluation of the relative binding affinity $\Delta\Delta A$. From $\Delta\Delta A$ we can determine which inhibitor has a greater binding affinity.[16]

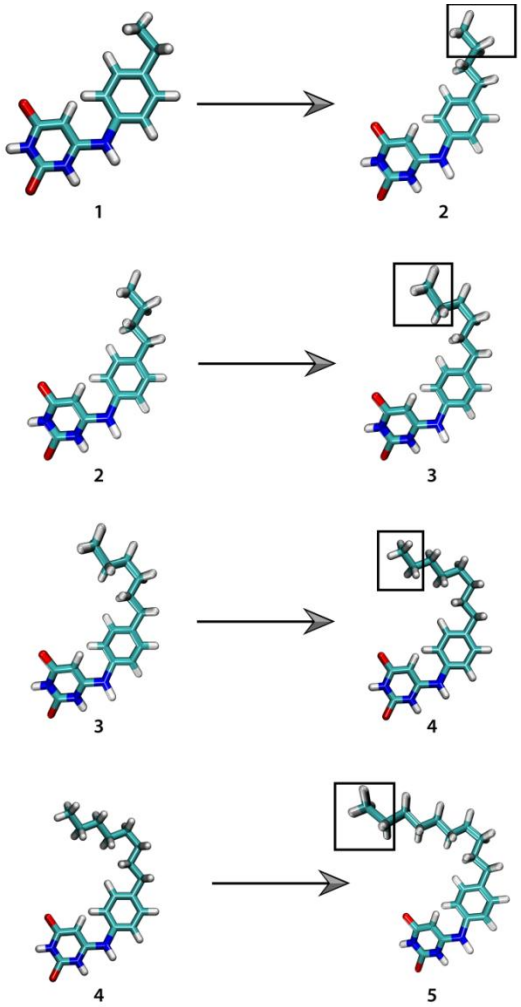


Figure 5.7 The 4 free energy perturbation simulations that were performed.

In each of the free energy perturbation simulations (Figure 5.7), a hydrogen in the reactant is replaced by a CH_2CH_3 in the product. All simulations were performed using conditions mentioned in section 5.3.2. The dual-topology method was used and the intermediate points between the physical endpoints (λ_A = 0 and λ_B = 1) were defined at coupling parameter (λ) intervals of 0.5. The largest physical change to the system occurs at the endpoint causing interactions in the system to change drastically. To overcome the endpoint problem, or improve the convergence of the simulations, the second and second last λ intervals were set at 0.025 instead of 0.5. The bond and angle term in the potential energy function were unperturbed in order to maintain the structure of the perturbed part of the system for λ values close to the endpoint values 0 or 1.

The starting coordinates for each of the simulations were obtained from the individual molecular dynamics simulations for each inhibitor. For each λ window, 700ps of equilibration was performed followed by 1.5ns of data collection. Double-wide sampling was used over the full range of λ to calculate the overall free energy difference in the transformations.[16]

5.6 Initial Preparation

Initial coordinates for both systems were obtained from the Brookhaven Protein Data Bank. The crystal structures used were resolved at 1.9Å and 1.75Å for 1SSP (hUDG)[17] and 1UDG (hsvUDG)[18] respectively. The protein was prepared (structure corrected[19] and protonation[20] states determined) in the manner as mentioned in chapter 4.3.2. Using flexible docking methods and hydrophobic analyses on the hsvUDG protein revealed two potential hydrophobic pockets (Figure 5.8) in which the alkyl chain of the inhibitors could interact.[21]

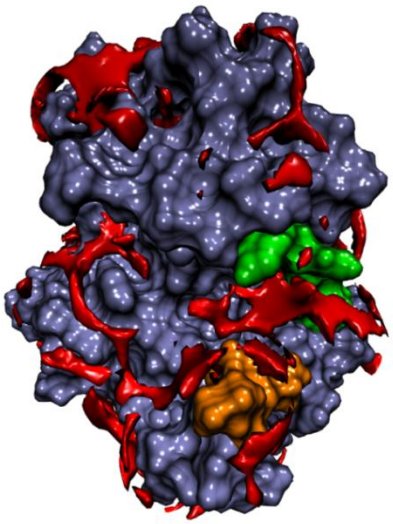


Figure 5.8 Two hydrophobic regions identified. Hydrophobic pocket 1 displayed in Green (VAL107, VAL103, GLN95, TYR90, PRO108 and PRO110) and hydrophobic pocket 2 displayed in Orange (SER212, PRO213, LEU214 and PRO111) represent the two regions. Red represents hydrophobic character.

2ns simulations were carried out using the 6-(4-octylanilino)-uracil docked in the hydrophobic pocket 1 and 2. It was determined that hydrophobic pocket 1 has a stronger interaction with the inhibitor than hydrophobic pocket 2. This

initial assumption was based on the duration with which the alkyl chain of the inhibitors remained in the binding pocket. When the alkyl chain was placed in hydrophobic pocket 2, it moved out of the pocket, whereas in hydrophobic pocket 1, it remained in the pocket. All simulations were carried out with the inhibitors docked in hydrophobic pocket 1.

5.7 Results and Discussion

5.7.1 Free Energy Perturbation Results

Tables 5.5 and 5.6 show the $\Delta\Delta A$ for the 6-(4-alkylanilino)-uracil inhibitors in hsvUDG and hUDG respectively. For both proteins, it can be seen that inhibitor 4 exhibits the strongest binding. All $\Delta\Delta A$ are negative, until the simulation transformation going from inhibitor 4 to inhibitor 5 in the hsvUDG enzyme (Table 5.5). Backward perturbation were performed and converged with negligible error. This indicates a decrease in the binding. This is in agreement with the IC_{50} based experimental results for the hsvUDG enzyme shown in Table 5.1 which ranks the inhibitors in the order of **4** > **3** > **5** > **2** > **1.** The fact that these results agree with the experimental findings confirms and validates the force field parameters and computational techniques used in this study.

Perturbation	ΔA_3(kcal.mol^{-1})	ΔA_4 (kcal.mol^{-1})	$\Delta\Delta A$ (kcal.mol^{-1})
1 --> **2**	-1.666	-2.570	-0.904
2 --> **3**	-0.263	-1.830	-1.567
3 --> **4**	0.228	-1.532	-1.760
4 --> **5**	-0.238	1.650	1.888

Table 5.5 Relative binding free energies of the inhibitors in hsvUDG.

Perturbation	ΔA_3(kcal.mol^{-1})	ΔA_4 (kcal.mol^{-1})	$\Delta\Delta A$ (kcal.mol^{-1})
1 --> **2**	-1.666	-5.699	-4.033
2 --> **3**	-0.263	-1.603	-1.340
3 --> **4**	0.228	-0.684	-0.912
4 --> **5**	-0.238	0.215	0.453

Table 5.6 Relative binding free energies of the inhibitors in hUDG.

5.7.2 Inhibitor Protein Interaction Profile Analyses

As mentioned in chapter 4.4.1, hydrogen bonds, van der Waals, electrostatic, water-mediated and π – stacking interactions are the most common types of interactions that occur between the protein and its substrate. All simulations were allowed to run for 10ns of production simulation. Two interaction profile figures are shown to display the way in which the position of the inhibitors vary as the alkyl chain increases in length. Figure 5.10(A) of the interaction profiles illustrates a flat 2D orientation of the inhibitor with all relevant amino acids around it. Figure 5.10(B) shows the 3D orientation of the inhibitor with respect to the amino acids that form the sides and base of the binding pocket. Starting points for each inhibitor were all based on the same position as the strongest inhibitor (inhibitor with the highest IC_{50} value). Giving them all the same starting position, which is quite deep inside the binding pocket, would eliminate discrepancies. Strong inhibitors would remain in the binding pocket, whereas weaker inhibitors would have weaker binding to the enzyme.

Electrostatic interactions and van der Waals interactions make up the nonbonded interaction energy that is calculated in each case. The binding pocket of both the hsvUDG and hUDG proteins can be decomposed into two regions of interest. “Region A” includes all amino acids that are directly in the binding pocket of the protein, and “Region B” includes all amino acids located just outside the binding pocket that interact with the benzene ring structure and the hydrophobic alkyl chain of the 6-(4-alkylanilino)-uracil inhibitors.

5.7.3 Rationalising Inhibitor Behaviour in the hsvUDG Enzyme

The average interaction energy between the inhibitor and important amino acids was calculated (Table 5.7) and the regions (A and B) of the protein can be seen in Figure 5.9. In hsvUDG, the ASN147 amino acid can be considered the deepest amino acid in the binding pocket. The binding affinity of the inhibitors can be qualitatively assessed by their average distance from the ASN147 amino acid.[18] The reason for including so many amino acids that constitute “Region B” is because of the high degree of freedom the alkyl chains possess. Certain inhibitors

do not bind strongly and their alkyl chain moves between the two hydrophobic pockets (Figure 5.8) and interact with many amino acids very weakly. However, others seem to bind strongly within hydrophobic pocket 1 and display very little movement.

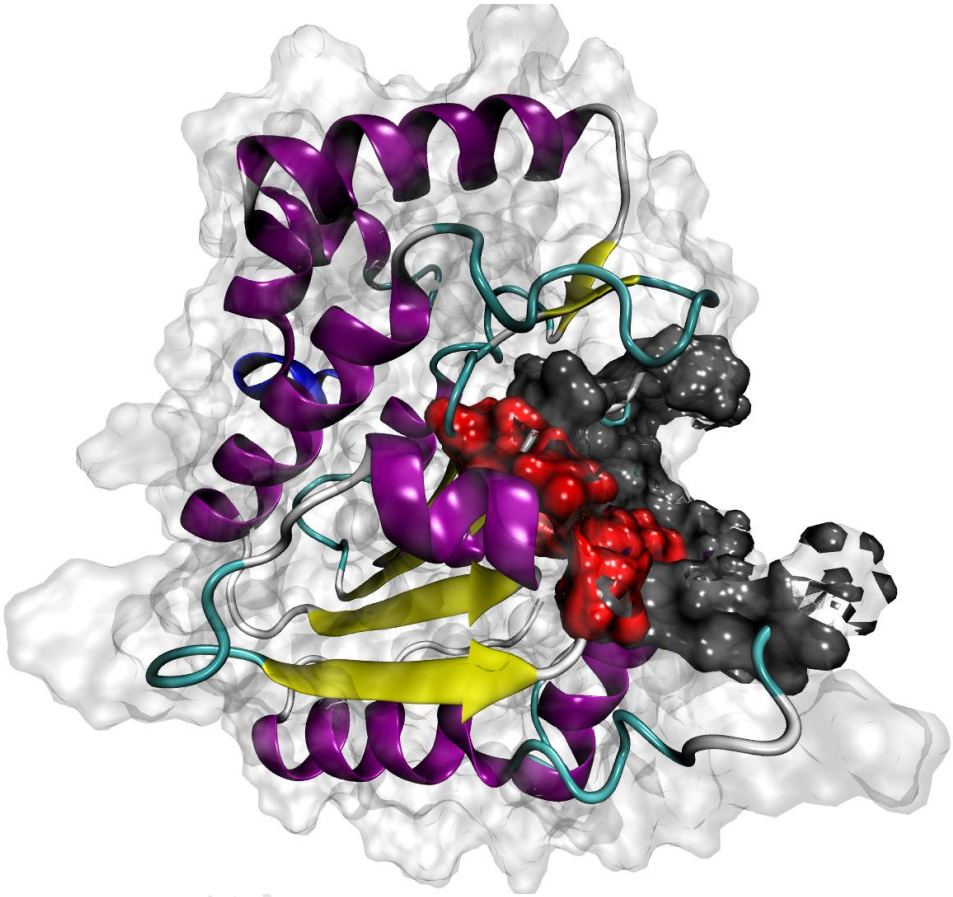


Figure 5.9 Illustration of the two regions of the hsvUDG enzyme, region A (red) and region B (grey) that are considered in the interaction profile for the inhibitors.

Average Interaction Energy (kcal.mol^{-1})																
REGION	AMINO ACID	1			2			3			4			5		
		TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC
A	PHE101	-7.42	-4.26	-3.16	-7.50	-3.97	-3.54	-7.02	-4.20	-2.82	-7.74	-4.37	-3.37	-7.60	-4.25	-3.36
	HIS210	-15.44	-0.27	-15.17	-16.73	-0.99	-15.74	-16.23	0.33	-16.56	-16.08	0.55	-16.63	-16.21	0.34	-16.54
	SER112	-5.30	-1.19	-4.11	-3.20	-1.52	-1.68	-1.43	-1.37	-0.06	-2.84	-1.55	-1.29	-2.68	-1.71	-0.97
	GLN87	-2.98	-0.41	-2.57	-3.75	-1.35	-2.40	-2.15	-0.61	-1.55	-2.21	-0.42	-1.79	-2.05	-0.51	-1.54
	ASN147	-9.91	-0.26	-9.66	-10.01	-0.36	-9.69	-9.76	-0.33	-9.42	-9.94	-0.36	-9.58	-9.60	-0.30	-9.39
	GLY86	-1.63	-0.82	-0.81	-1.62	-0.85	-0.77	-1.56	-0.80	-0.76	-1.55	-0.78	-0.76	-1.63	-0.77	-0.86
	ALA100	-2.49	-1.20	-1.29	-2.55	-1.11	-1.44	-2.51	-1.26	-1.26	-2.55	-1.20	-1.35	-2.48	-1.20	-1.28
	TOTAL A	**-45.20**	**-8.43**	**-36.77**	**-45.39**	**-10.1513**	**-35.24**	**-40.66**	**-8.23**	**-32.43**	**-42.93**	**-8.14**	**-34.79**	**-42.35**	**-8.40**	**-33.94**
B	TYR90	-6.67	-4.88	-1.80	-7.19	-5.20	-1.20	-8.36	-0.33	-1.55	-9.86	-8.34	-1.53	-8.95	-7.63	-1.32
	HIS92	-0.05	-0.27	0.013	-0.20	-0.19	-0.01	-1.08	-1.04	-0.04	-2.31	-2.27	-0.04	-2.17	-2.15	-0.02
	PRO110	-0.56	-0.24	-0.33	-0.27	-0.24	-0.03	-1.28	-0.96	-0.32	-2.44	-2.07	-0.37	-2.74	-2.37	-0.37
	VAL103	-0.12	-0.05	-0.07	-0.12	-0.05	-0.07	-0.23	-0.15	-0.08	-0.84	-0.73	-0.11	-0.63	-0.53	-0.10
	HIS91	-0.90	-0.89	-0.01	-0.96	-0.84	-0.12	-2.63	-2.54	-0.09	-3.02	-2.88	-0.14	-2.22	-2.15	-0.07
	PRO108	0	0	0	0	0	0	-0.07	-0.039	-0.03	-0.38	-0.30	-0.08	-0.72	-0.70	-0.03
	PRO111	-0.37	-0.20	-0.17	-0.24	-0.22	-0.02	-0.50	-0.37	-0.13	-0.50	-0.43	-0.07	-0.70	-0.62	-0.08
	PRO213	-1.05	-1.05	0	-1.09	-1.09	0	-0.60	-0.58	-0.02	-0.39	-0.36	-0.03	-0.77	-0.75	-0.02
	LEU214	-0.43	-0.35	-0.08	-0.35	-0.32	-0.03	-0.19	-0.14	-0.05	0.03	-0.10	-0.07	-0.10	-0.06	-0.04
	ASP88	-5.51	-2.58	-2.93	-6.11	-3.11	-3.00	-6.59	-2.83	-3.76	-6.42	-2.72	-3.71	-6.25	-2.58	-3.67
	PRO89	-3.52	-2.42	-1.1	-2.78	-2.63	-0.15	-4.22	-2.70	-1.52	-4.24	-2.61	-1.63	-3.94	-2.57	-1.37
	GLN95	-0.52	-0.14	-0.38	-0.31	-0.16	-0.15	-0.70	-0.57	-0.13	-1.48	-1.38	-0.10	-1.70	-1.58	-0.12
	VAL107	0	0	0	0	0	0	-0.06	-0.05	-0.01	-0.77	-0.74	-0.03	-1.01	-0.99	-0.02
	PRO211	-2.13	-1.05	-1.08	-1.00	-0.94	-0.06	-1.62	-0.54	-1.08	-0.91	-0.61	-0.30	-1.32	-0.63	-0.69
	SER212	-0.85	-0.78	-0.07	-0.81	-0.78	-0.02	-0.33	-0.28	-0.05	-1.02	-0.45	-0.57	-0.23	-0.22	-0.01
	TOTAL B	**-23.00**	**-14.89**	**-7.19**	**-18.20**	**-12.67**	**-5.03**	**-21.88**	**-13.12**	**-8.56**	**-33.08**	**-25.95**	**-6.90**	**-33.30**	**-25.53**	**-7.27**
	TOTAL PROTEIN	**-67.62**	**-25.07**	**-42.55**	**-67.92**	**-28.99**	**-38.93**	**-69.32**	**-30.26**	**-39.05**	**-78.05**	**-36.85**	**-41.20**	**-76.54**	**-36.96**	**-36.96**

Table 5.7 Average interaction energies between inhibitors 1-5, and key amino acid residues in hsvUDG.

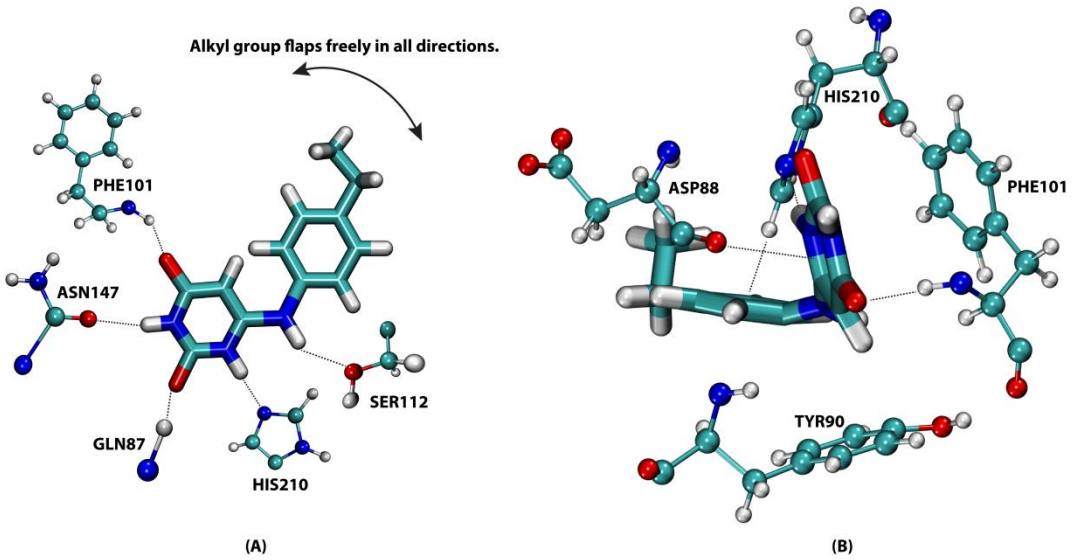


Figure 5.10 (A) 2-dimensional flat view and (B) 3-dimensional spatial view of the inhibitor 1 interactions profile with hsvUDG.

With an average interaction energy of -67.62kcal.mol^{-1} the uracil head of inhibitor 1 forms 4 hydrogen bonds with the protein. The carboxyl group of ASN147 and the nitrogen from the HIS210 ring accepts hydrogen bonds from the hydrogen on N3 and N1 of the uracil head respectively. PHE101 and GLN87 donate hydrogen bonds to O4 and O2 on the uracil head respectively (Figure 5.10 (A)). The alkyl tail moves freely in all directions. Figure 5.11 (A) and (B) show the interaction energy and distance time series of inhibitor 1 with ASN147 and the hsvUDG enzyme.

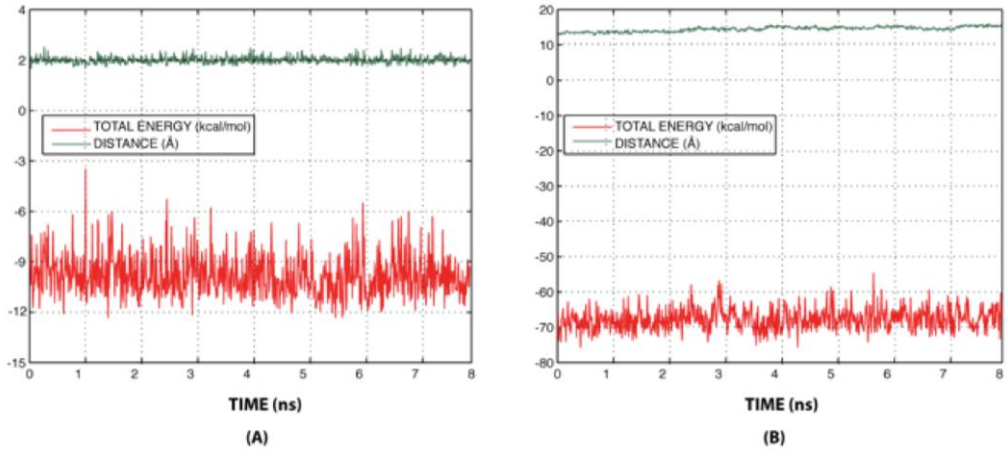


Figure 5.11 The interaction energy (red) and the distance (green) time series plot for inhibitor 1 with, (A) ASN147 and (B) the complete hsvUDG enzyme.

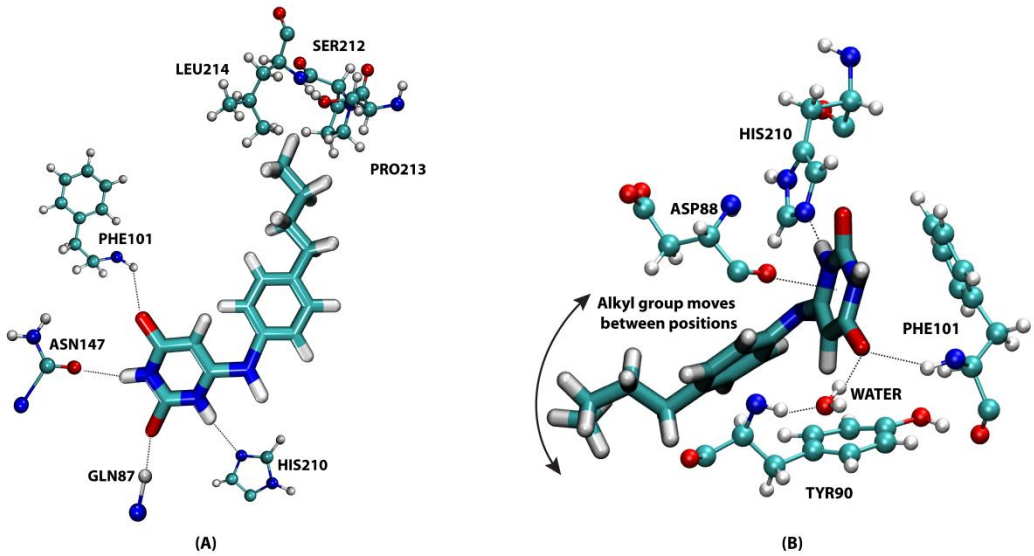


Figure 5.12 (A) 2-dimensional flat view and (B) 3-dimensional spatial view of the inhibitor 2 interactions profile with hsvUDG.

Inhibitor 2 has a similar interaction profile as inhibitor 1 (Figure 5.12 (A)). However, a consistent water-mediated bond between TYR90 and O4 of the uracil head is observed (Figure 5.12 (B)). Inhibitor 2 produces an average interaction energy of -67.92kcal.mol^{-1} with hsvUDG. The alkyl chain tail moves freely between hydrophobic pockets 1 and 2 shown previously in Figure 5.8. Figure 5.13 (A) and (B) show the interaction energy and distance time series of inhibitor 2 with ASN147 and the hsvUDG enzyme respectively.

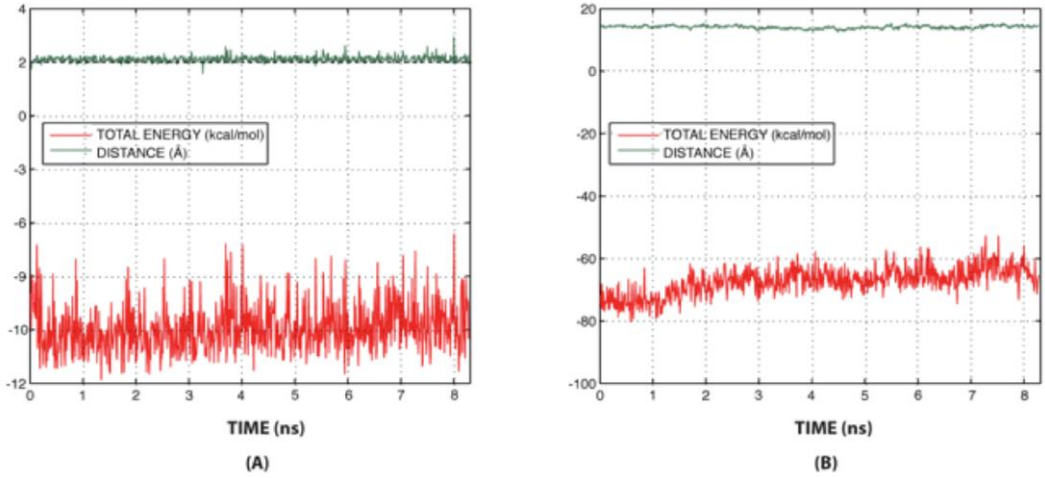


Figure 5.13 The interaction energy (red) and the distance (green) time series plot for inhibitor 2 with, (A) ASN147 and (B) the complete hsvUDG enzyme.

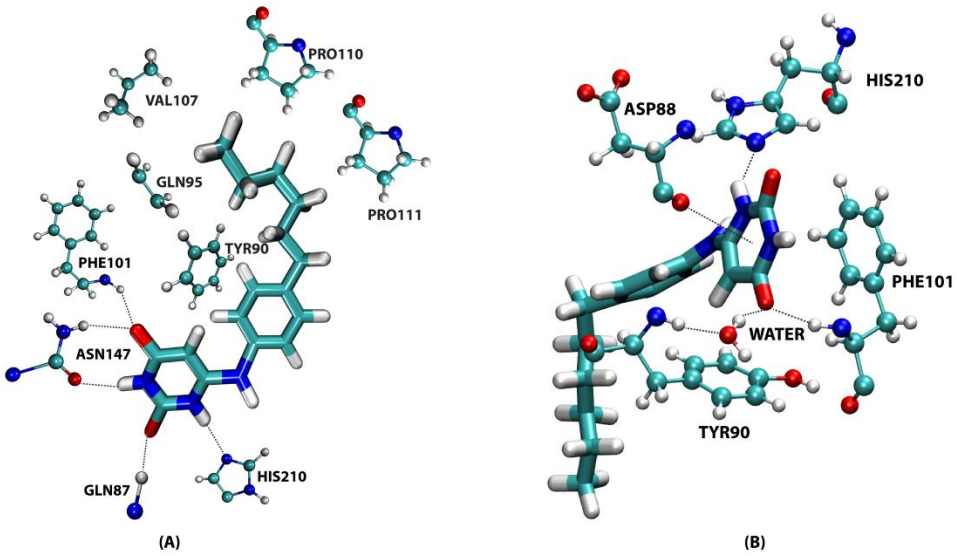


Figure 5.14 (A) 2-dimensional flat view and (B) 3-dimensional spatial view of the inhibitor 3 interactions profile with hsvUDG.

Inhibitor 3 possesses a bidentate hydrogen bond interaction with the hydrogen of N3 and the oxygen of C4 on uracil. Inhibitor 3 produces an average interaction energy of -69.32kcal.mol^{-1} with hsvUDG. The water-mediated bond between TYR90 and O3 is present in inhibitor 3 just as it is present for inhibitor 2 (Figure 5.14 (B)). However in Figure 5.14 (A), the alkyl chain remains in hydrophobic pocket 1 wrapping itself around TYR90. Figure 5.15 (A) and (B) show the interaction energy and distance time series of inhibitor 3 with ASN147 and the hsvUDG enzyme respectively.

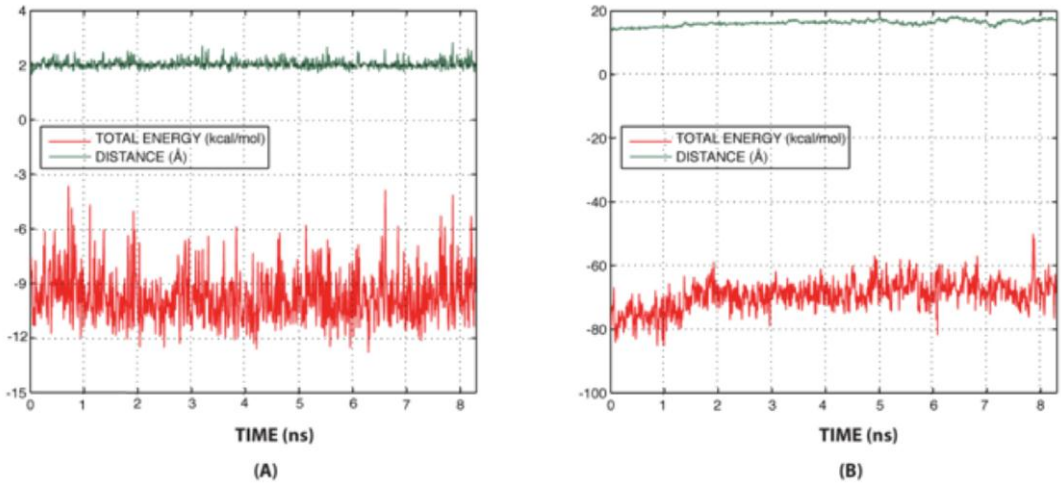


Figure 5.15 The interaction energy (red) and the distance (green) time series plot for inhibitor 3 with, (A) ASN147 and (B) the complete hsvUDG enzyme.

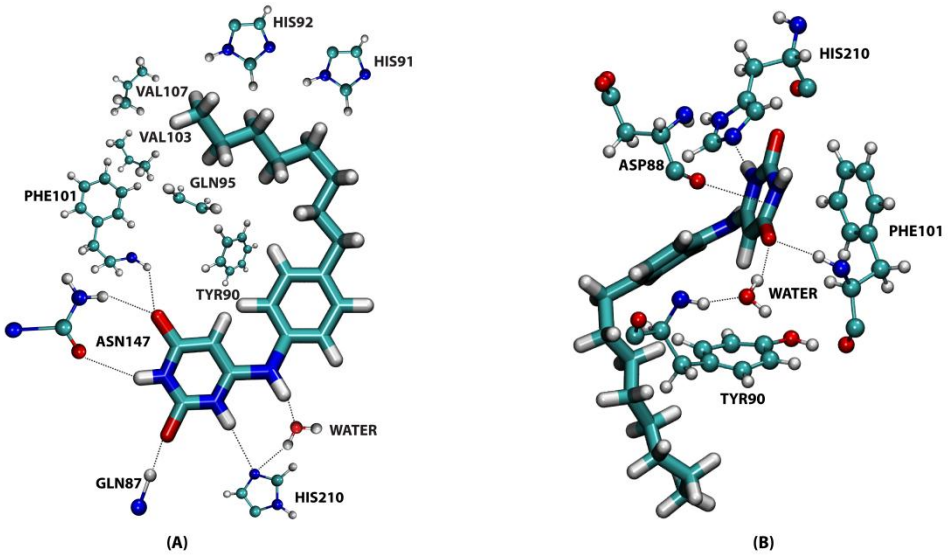


Figure 5.16 (A) 2-dimensional flat view and (B) 3-dimensional spatial view of the inhibitor 4 interactions profile with hsvUDG.

Inhibitor 4 has an average interaction energy of -78.05kcal.mol^{-1} with the hsvUDG enzyme, the greatest binding affinity of all the inhibitors (Figure 5.16). This is in agreement with experimental values. Figure 5.17 (A) and (B) show the interaction energy and distance time series of inhibitor 4 with ASN147 and the hsvUDG enzyme respectively.

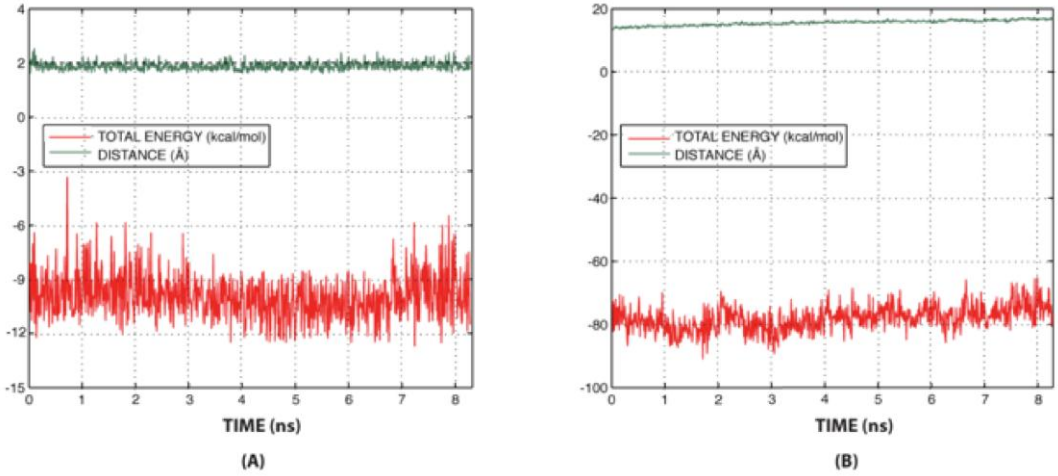


Figure 5.17 The interaction energy (red) and the distance (green) time series plot for inhibitor 4 with, (A) ASN147 and (B) the complete hsvUDG enzyme.

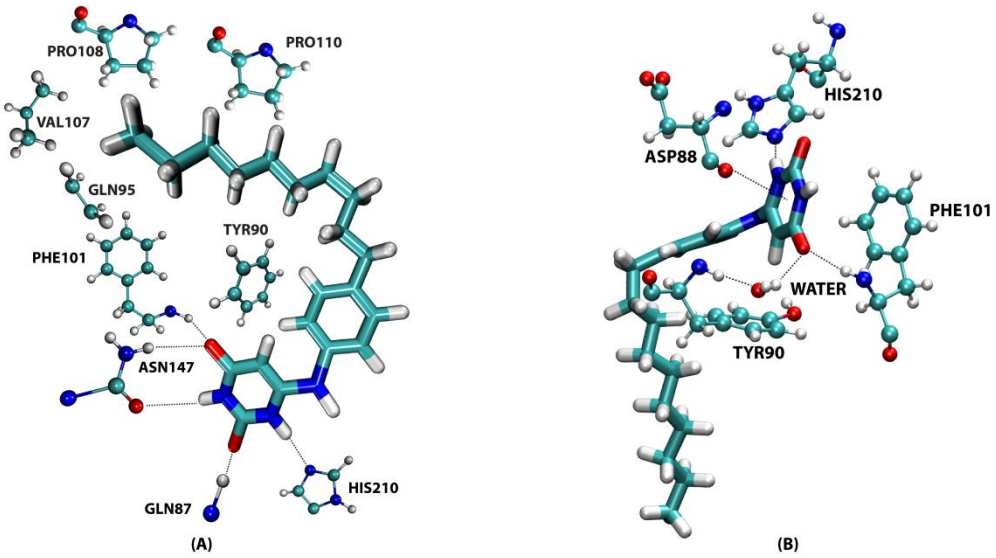


Figure 5.18 (A) 2-dimensional flat view and (B) 3-dimensional spatial view of the inhibitor 5 interactions profile with hsvUDG.

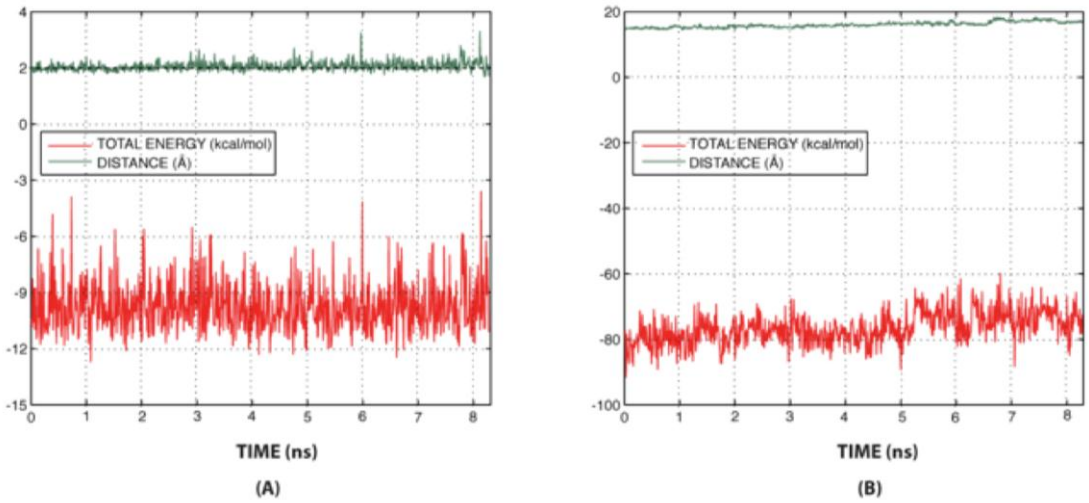


Figure 5.19 The interaction energy (red) and the distance (green) time series plot for inhibitor 5 with, (A) ASN147 and (B) the complete hsvUDG enzyme.

Figure 5.19 (A) and (B) show the interaction energy and distance time series of inhibitor 5 with ASN147 and the hsvUDG enzyme respectively. Inhibitors 3, 4 and 5 possess the ability to have their alkyl chains form a favourable interaction with hydrophobic pocket 1, formed by VAL107, VAL103, GLN95, TYR90, PRO108 and PRO110. The alkyl chain for inhibitors 3, 4 and 5 remains in hydrophobic pocket 1 for the duration of the simulations. All the inhibitors form a water-

mediated interaction with TYR90, excluding inhibitor 1, which does not penetrate the binding pocket deep enough.

5.7.4 Rationalising Inhibitor Behaviour in the hUDG Enzyme

The average interaction energy between the inhibitor and important amino acids was calculated (Table 5.8) and the regions (A and B) of the protein can be seen in Figure 5.20. In hUDG, ASN204 can be considered the deepest amino acid in the binding pocket as it interacts with the uracil head of the natural uracil substrate in DNA. Therefore, the farther the inhibitor is on average from ASN204, the less likely the inhibitor is to be favoured within the binding pocket.[17] This is considered to be a rough estimate of the binding capability of the inhibitor. Due to the highly conserved nature of the family of uracil-DNA glycosylase enzymes (Figure 5.1), very similar interaction as seen in hsvUDG, will be seen in hUDG.

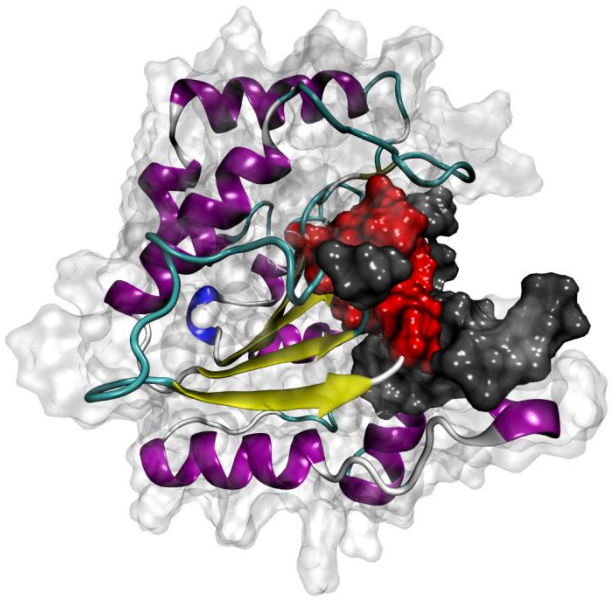


Figure 5.20 Illustration of the two regions of the hUDG enzyme, region A (red) and region B (grey) that are considered in the interaction profile for the inhibitors.

Average Interaction Energy (kcal.mol^{-1})																
REGION	AMINO ACID	1			2			3			4			5		
		TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC
A	ASN204	-0.01	-0.01	0	-9.65	-0.32	-9.33	-9.11	-0.60	-8.51	-8.76	-0.45	-8.31	-8.53	-0.39	-8.14
	CYS157	-0.46	-0.39	-0.07	-2.27	-1.45	-0.83	-3.26	-1.99	-1.28	-2.20	-1.52	-0.68	-2.34	-1.43	-0.91
	ASP145	-3.98	-2.38	-1.60	-6.33	-2.75	-3.58	-13.37	-2.96	-10.40	-7.12	-2.48	-4.64	-9.31	-3.20	-6.12
	PHE158	-0.89	-0.71	-0.18	-5.91	-3.50	-2.41	-6.19	-3.18	-3.01	-7.03	-4.14	-2.89	-7.19	-4.04	-3.15
	TYR147	-6.78	-5.22	-1.56	-7.24	-5.15	-2.10	-7.02	-3.12	-3.90	-7.67	-5.51	-2.16	-6.83	-4.79	-2.04
	GLN144	-1.30	-0.79	-0.51	-3.15	-0.77	-2.38	-2.32	-0.61	-1.71	-2.77	-0.67	-2.09	-2.86	-0.73	-2.12
	GLY143	-0.47	-0.18	-0.29	-1.84	-0.76	-1.08	-1.57	-0.77	-0.80	-1.77	-0.73	-1.04	-1.53	-0.89	-0.65
	HIS268	-2.03	-0.79	-1.24	-16.10	-0.94	-15.15	-3.89	-0.43	-3.47	-1.42	-1.39	-0.03	-1.45	-1.95	0.50
	PRO146	-4.30	-1.50	-2.79	-3.86	-2.78	-1.09	-3.45	-1.72	-1.73	-3.67	-2.51	-1.16	-4.54	-2.38	-2.16
	TOTAL A	**-20.29**	**-12.04**	**-8.25**	**-56.36**	**-18.42**	**-37.94**	**-50.19**	**-15.38**	**-34.80**	**-42.41**	**-19.41**	**-23.00**	**-44.59**	**-19.81**	**-24.78**
B	PRO269	-0.92	-0.65	-0.27	-1.35	-1.55	0.20	-3.30	-2.91	-0.39	-1.15	-1.08	-0.07	-2.42	-2.28	-0.14
	PRO168	-0.31	-0.26	-0.05	-0.31	-0.18	-0.12	-0.55	-0.50	-0.05	-0.93	-0.82	-0.11	-1.30	-1.16	-0.13
	PRO167	-0.86	-0.86	0	-0.57	-0.36	-0.21	-0.29	-0.21	-0.08	-0.92	-0.78	-0.14	-0.65	-0.56	-0.09
	HIS148	-5.23	-3.45	-1.78	-2.24	-2.12	-0.12	-0.24	-0.15	-0.09	-1.54	-1.51	-0.03	-1.18	-1.00	-0.18
	SER169	-1.55	-1.34	-0.22	-2.87	-1.44	-1.43	-4.10	-2.73	-1.37	-3.79	-2.24	-2.24	-4.33	-2.14	-2.18
	SER247	0	0	0	-0.01	-0.01	0	-0.01	-0.01	0	-0.20	-0.20	-0.01	0	0	0
	ILE173	-0.10	-0.10	0	-0.52	-0.37	-0.15	-1.44	-1.35	-0.09	-0.07	-0.05	-0.02	-0.34	-0.32	-0.02
	SER270	-0.45	-0.31	-0.14	-0.50	-0.46	-0.04	-1.52	-1.40	-0.12	-0.84	-0.83	-0.01	-1.54	-1.52	-0.02
	LEU272	-0.02	-0.03	0.01	0	0	0	-0.14	-0.13	-0.01	-0.75	-0.71	-0.03	-0.48	-0.43	-0.05
	SER273	0	0	0	0	0	0	-0.19	-0.13	-0.06	-0.07	-0.05	-0.02	0	0	0
	LEU170	-0.33	-0.34	0.01	-0.55	-0.41	-0.14	-1.04	-1.00	-0.04	-0.80	-0.73	-0.07	-0.46	-0.37	-0.09
	ASN172	0	0	0	0	0	0	-1.43	-1.23	-0.20	-0.39	-0.49	-0.49	-0.38	-0.32	-0.06
	PRO271	-0.31	-0.29	-0.02	-0.19	-0.21	0.02	-2.08	-2.07	-0.01	-0.98	-0.97	-0.01	-1.13	-1.13	0
	GLN152	-1.10	-0.73	-0.37	0	0	0	0	0	0	-0.51	-0.46	-0.05	-0.31	-0.30	-0.01
	TOTAL B	**-11.09**	**-8.39**	**-2.70**	**-8.94**	**-7.33**	**-1.61**	**-16.74**	**-13.84**	**-2.90**	**-13.88**	**-10.93**	**-2.86**	**-14.50**	**-11.62**	**-2.30**
	TOTAL PROTEIN	**-36.80**	**-22.08**	**-14.72**	**-65.65**	**-28.36**	**-37.29**	**-69.08**	**-33.45**	**-35.63**	**-66.78**	**-34.65**	**-32.13**	**-59.99**	**-34.68**	**-25.31**

Table 5.8 Average interaction energies between inhibitors 1-5, and key amino acid residues in hUDG.

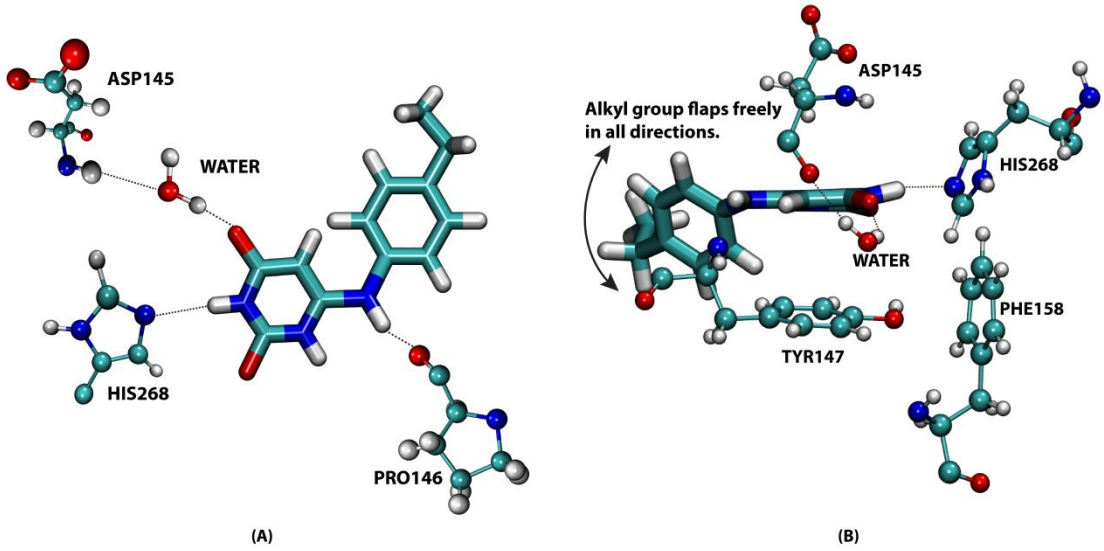


Figure 5.21 (A) 2-dimensional flat view and (B) 3-dimensional spatial view of the inhibitor 1 interaction profile with hUDG.

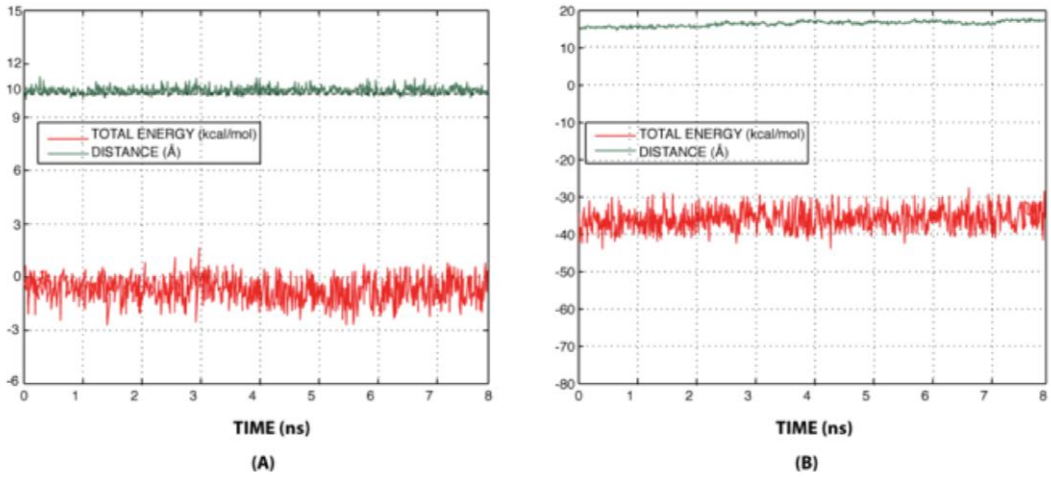


Figure 5.22 The interaction energy (red) and the distance (green) time series plot for inhibitor 1 with, (A) ASN204 and (B) the complete hUDG enzyme.

Inhibitor 1 has a low binding affinity for hUDG and does not make significant contacts with the catalytic residues. It donates two hydrogen bonds to HIS268 and PRO146 as shown in Figure 5.21. The average interaction energy between inhibitor 1 and hUDG is -36.81kcal.mol^{-1}. This is the lowest interaction energy of all the inhibitors. A water-mediated interaction forms between the O4 atom of uracil and ASP145. Figure 5.22 (A) and (B) show the interaction energy and

distance time series of inhibitor 1 with ASN204 and the hUDG enzyme respectively.

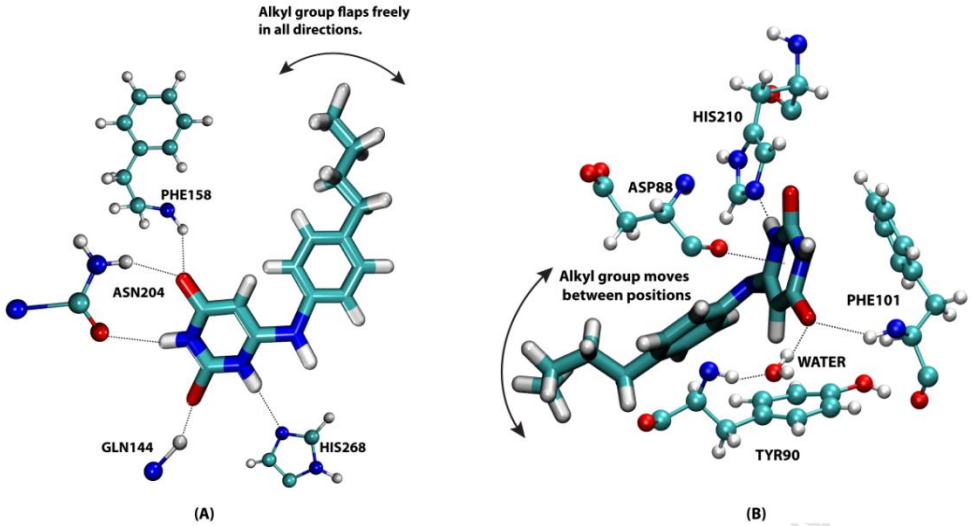


Figure 5.23 (A) 2-dimensional flat view and (B) 3-dimensional spatial view of the inhibitor 2 interactions profile with hUDG.

Inhibitor 2 penetrates the binding pocket more deeply than inhibitor 1 and forms a bidentate interaction with the carboxamide functional group of ASN204. (Figure 5.23). The alkyl group in inhibitor 1 and 2 are not fixed in any hydrophobic pocket and therefore have many degrees of freedom. A water-mediated interaction forms between TYR90 and the uracil head of inhibitor 2 (Figure 5.23 (B)). Figure 5.24 (A) and (B) show the interaction energy and distance time series of inhibitor 2 with ASN204 and the hUDG enzyme respectively.

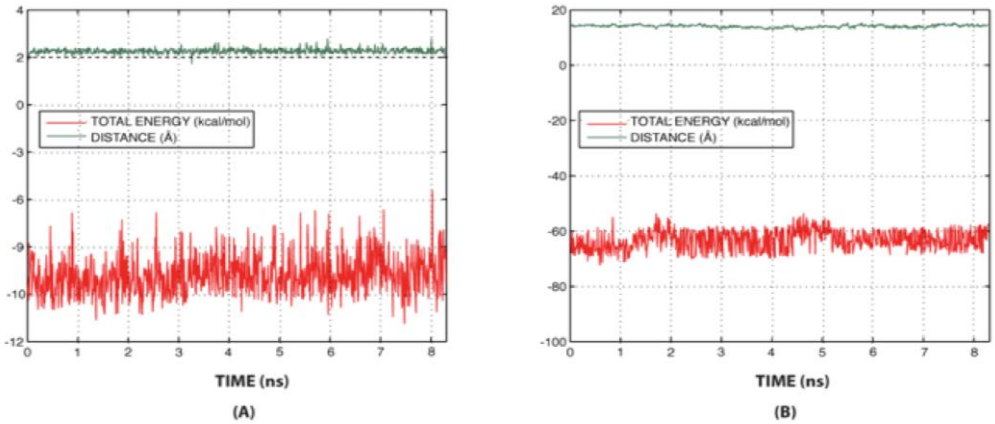


Figure 5.24 The interaction energy (red) and the distance (green) time series plot for inhibitor 2 with, (A) ASN204 and (B) the complete hUDG enzyme.

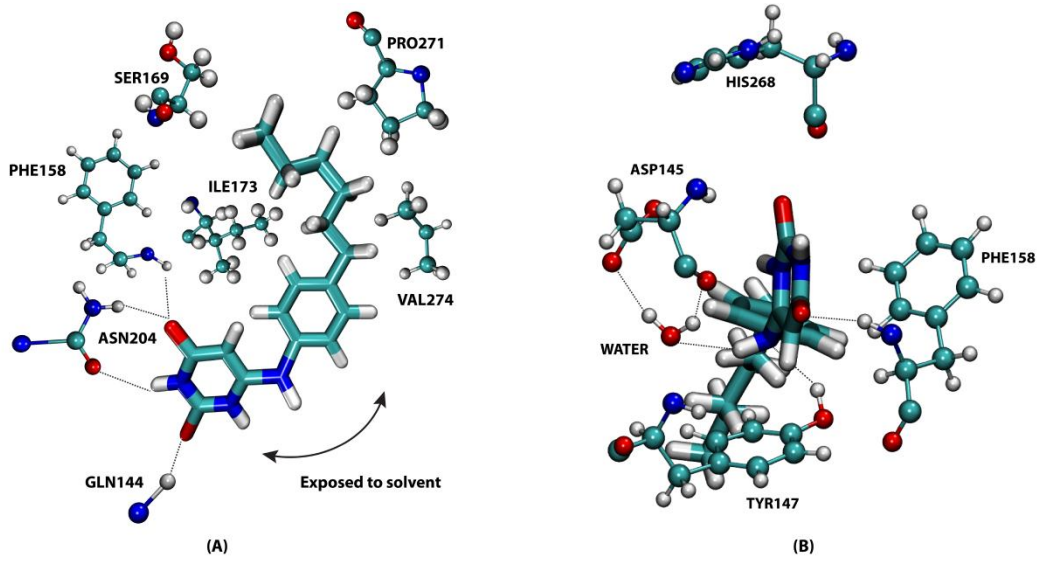


Figure 5.25 (A) 2-dimensional flat view and (B) 3-dimensional spatial view of the inhibitor 3 interactions profile with hUDG.

Inhibitor 3 has an average interaction energy of -69.08kcal.mol^{-1} with the hUDG enzyme and forms a bidentate interaction with the carboxamide functional group of ASN204 (Figure 5.25). Figure 5.26 (A) and (B) show the interaction energy and distance time series of inhibitor 3 with ASN204 and the hUDG enzyme respectively.

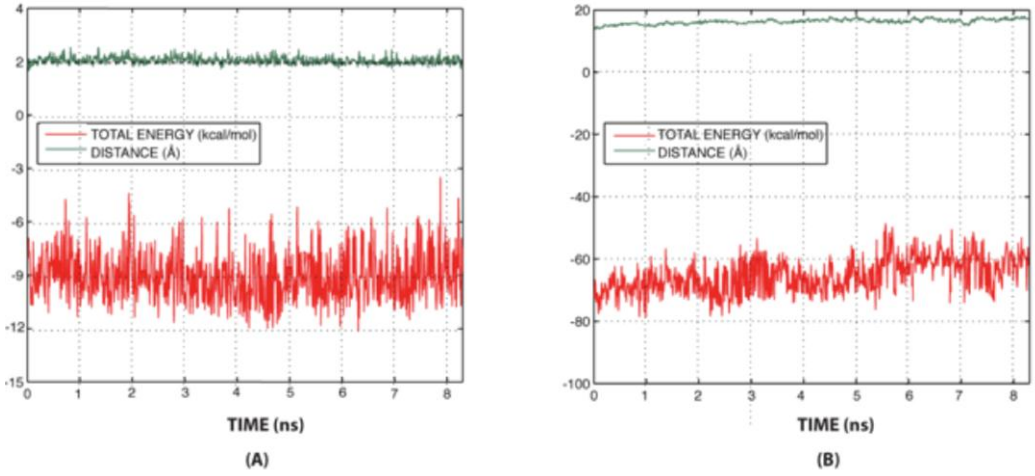


Figure 5.26 The interaction energy (red) and the distance (green) time series plot for inhibitor 3 with, (A) ASN204 and (B) the complete hUDG enzyme.

The most preferred hydrophobic pocket for the alkyl chain of inhibitors 3 (Figure 5.25), 4 (Figure 5.27) and 5 (Figure 5.28) is the SER270, PRO269 and

PRO271 pocket. This is equivalent to the SER212, PRO213 and PRO214 amino acids which form hydrophobic pocket 2 in hsvUDG.

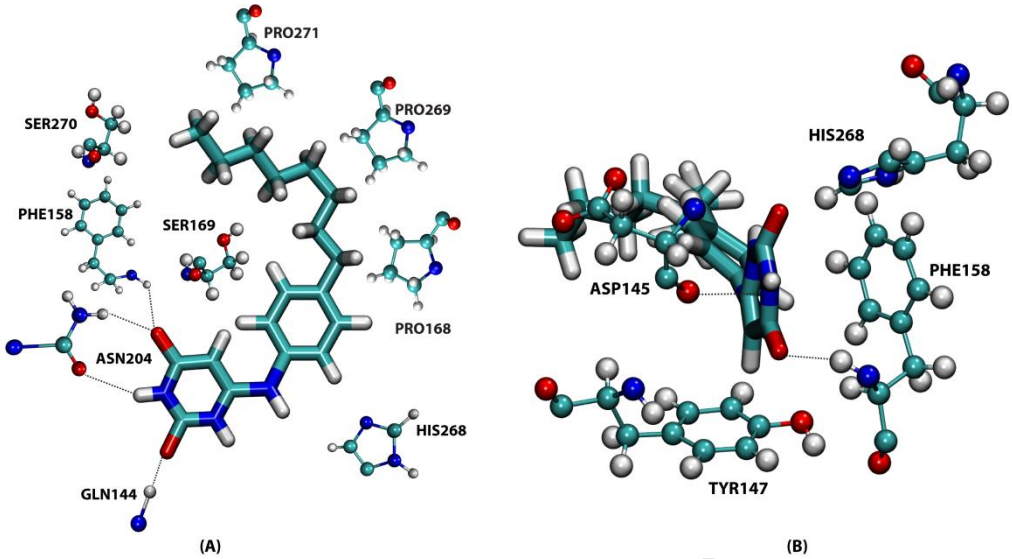


Figure 5.27 (A) 2-dimensional flat view and (B) 3-dimensional spatial view of the inhibitor 4 interactions profile with hUDG.

Figure 5.28 (A) and (B) show the interaction energy and distance time series of inhibitor 4 with ASN204 and the hUDG enzyme respectively.

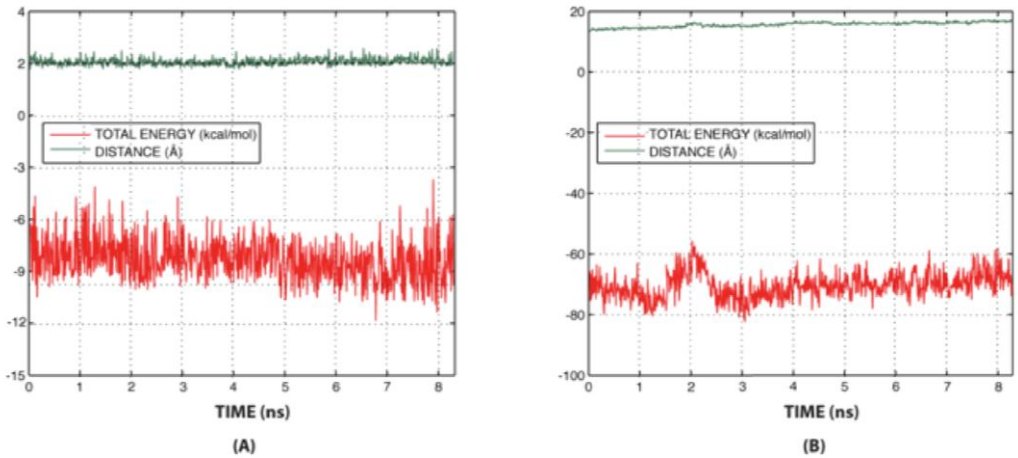


Figure 5.28 The interaction energy (red) and the distance (green) time series plot for inhibitor 4 with, (A) ASN204 and (B) the complete hUDG enzyme.

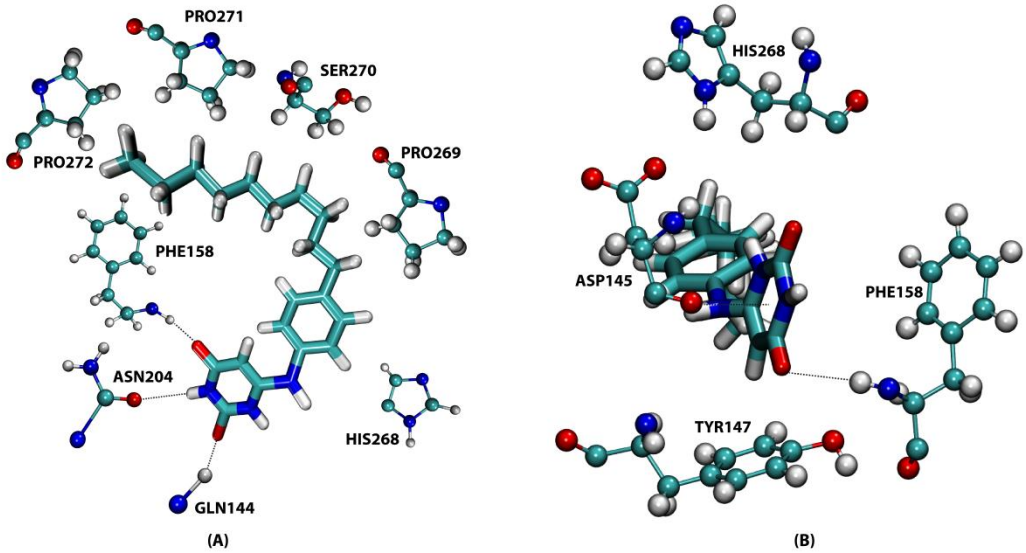


Figure 5.29 (A) 2-dimensional flat view and (B) 3-dimensional spatial view of the inhibitor 5 interactions profile with hUDG.

Figure 5.30 (A) and (B) show the interaction energy and distance time series of inhibitor 5 with ASN204 and the hUDG enzyme respectively.

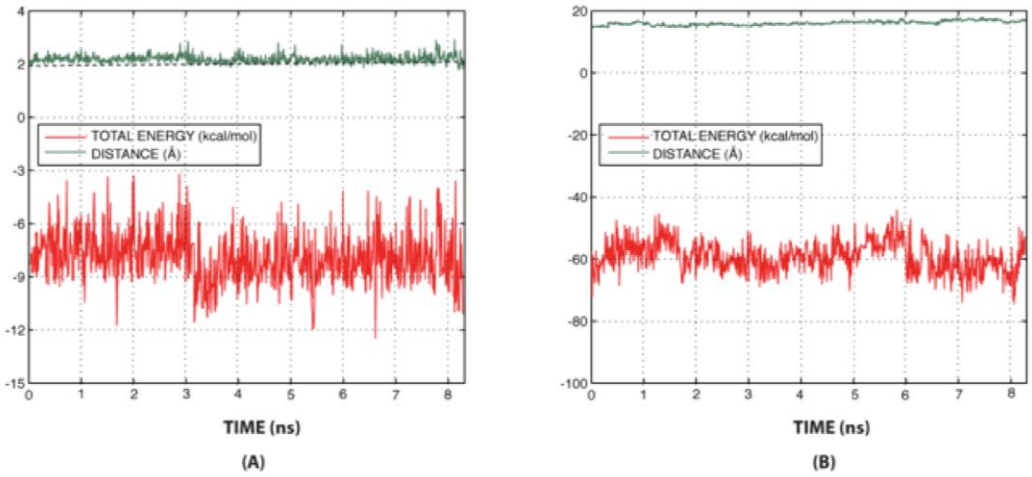


Figure 5.30 The interaction energy (red) and the distance (green) time series plot for inhibitor 5 with, (A) ASN204 and (B) the complete hUDG enzyme.

5.8 Overall Rationalisation of Inhibitor Binding Behaviour

For the hsvUDG protein, it can be seen that hydrophobic pocket 1 shown in Figure 5.8 plays an important role in explaining the trend in the inhibition effect of the inhibitors. As the alkyl chain length is increased from 2 carbons to 8 carbons, we see an increase in the interaction energy of the inhibitors.

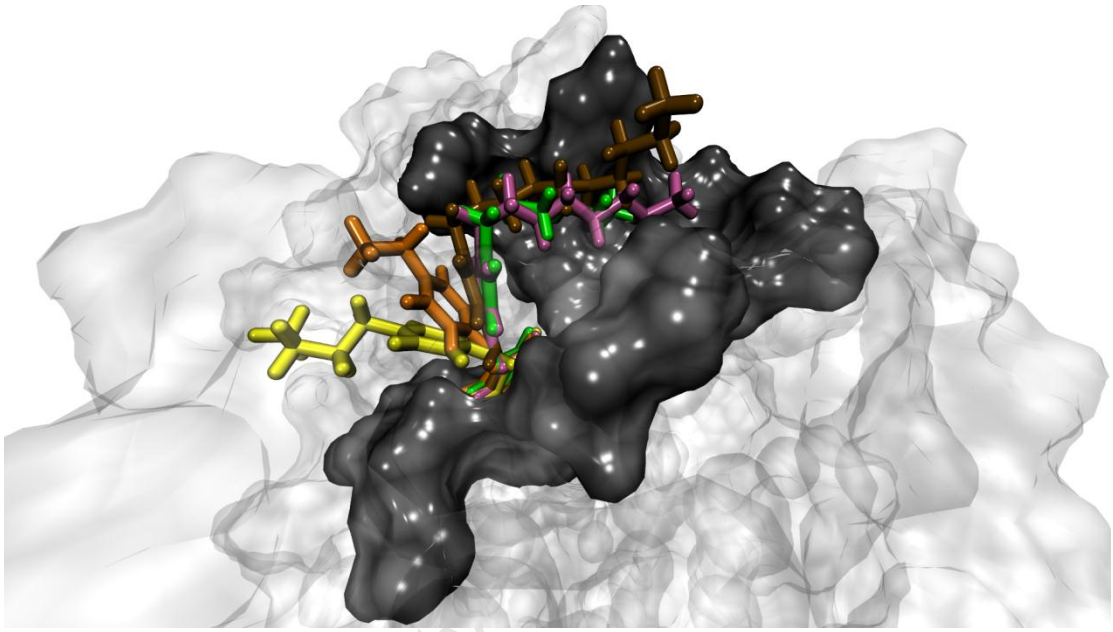


Figure 5.31 Overall comparison of inhibitor binding in hsvUDG. Orange = inhibitor 1, Yellow = Inhibitor 2, Green = Inhibitor 3, Pink = Inhibitor 4 and Brown = Inhibitor 5. The dark grey surface of the protein represents the amino acids forming the hydrophobic pocket.

When the transformation in the alkyl chain is increased from 8 carbons to 10 carbon atoms, a decrease in the inhibition effect is observed. From Figure 5.31, it can be seen that the alkyl chain of inhibitor 5 (Brown) is slightly too long to fit comfortably in the hydrophobic pocket. The binding affinity of the inhibitors seems to be dependent on the length of the alkyl chain.

Alkyl chains in the simulations of the inhibitors in the hUDG enzyme displayed a preference for hydrophobic pocket 2. Interaction energies between the inhibitor and the hsvUDG enzyme are on average 12.6kcal.mol^{-1} greater than the interaction energies of the inhibitors in hUDG.

Summary of Inhibitor Data in hsvUDG							
Inhibitors	Average Volume of Binding pocket ($Å^3$)	Protein Interaction Energy ($Kcal.mol^{-1}$)	IC_{50} Value (μM)	Nomalised ΔΔA values ($kcal.mol^{-1}$)	Distance From Binding-Pocket/ASN204 (Å)	RMSD of alkyl tail (Å)	Preferred hydrophobic pocket
1	289	-67.62	500	0	2.83	4.2	2
2	294	-67.92	150	-0.904	2.46	4.3	2
3	293	-69.32	30	-2.471	2.21	2.3	1
4	296	-78.05	8	-4.231	1.85	1.6	1
5	298	-76.54	35	-2.343	2.51	2.1	1

Table 5.9 Data summary for the hsvUDG protein.

Summary of Inhibitor Data in hUDG							
Inhibitors	Average Volume of Binding pocket ($Å^3$)	Protein Interaction Energy ($Kcal.mol^{-1}$)	IC_{50} Value (μM)	Normalised ΔΔA values ($kcal.mol^{-1}$)	Distance From Binding-Pocket/ASN204 (Å)	RMSD of alkyl tail (Å)	Preferred hydrophobic pocket
1	265	-36.80	>500	0	10.42	4.4	2
2	269	-65.65	>500	-4.032	2.4	4.3	2
3	274	-69.08	>300	-5.372	2.03	3.6	2
4	272	-66.78	>300	-6.284	2.05	3.3	2
5	275	-59.99	>500	-5.831	3.8	3.8	2

Table 5.10 Data summary for the hUDG protein.

From the RMSD of the alkyl tail in Table 5.9, it can be seen that if the tail is not long enough to enter hydrophobic pocket 1 in the hsvUDG protein, it is free to move around more. In Table 5.10, although the alkyl tail enters hydrophobic pocket 2 in hUDG, the RMSD indicates that the alkyl tail moves around quite a bit. On average the volume of the binding pocket in hUDG is approximately 20$Å^3$ smaller than the volume of the binding pocket in hsvUDG.

(A)

HYDROPHOBIC POCKET	AMINO ACID	Average Interaction Energy (kcal.mol^{-1})														
		1			2			3			4			5		
		TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC
1	TYR90	-6.67	-4.88	-1.80	-7.19	-5.20	-1.20	-8.36	-6.81	-1.55	-9.86	-8.34	-1.53	-8.95	-7.63	-1.32
	HIS92	-0.05	-0.27	0.013	-0.20	-0.19	-0.01	-1.08	-1.04	-0.04	-2.31	-2.27	-0.04	-2.17	-2.15	-0.02
	PRO110	-0.56	-0.24	-0.33	-0.27	-0.24	-0.03	-1.28	-0.96	-0.32	-2.44	-2.07	-0.37	-2.74	-2.37	-0.37
	VAL103	-0.12	-0.05	-0.07	-0.12	-0.05	-0.07	-0.23	-0.15	-0.08	-0.84	-0.73	-0.11	-0.63	-0.53	-0.10
	VAL107	0	0	0	0	0	0	-0.06	-0.05	-0.01	-0.77	-0.74	-0.03	-1.01	-0.99	-0.02
	GLN95	-0.52	-0.14	-0.38	-0.31	-0.16	-0.15	-0.70	-0.57	-0.13	-1.48	-1.38	-0.10	-1.70	-1.58	-0.12
	HIS91	-0.90	-0.89	-0.01	-0.96	-0.84	-0.12	-2.63	-2.54	-0.09	-3.02	-2.88	-0.14	-1.22	-1.15	-0.07
	PRO108	0	0	0	0	0	0	-0.07	-0.039	-0.03	-0.38	-0.30	-0.08	-0.72	-0.70	-0.03
	TOTAL	**-9.07**	**-6.47**	**-2.60**	**-9.05**	**-6.68**	**-2.21**	**-14.41**	**-12.16**	**-2.25**	**-21.11**	**-18.71**	**-2.40**	**-19.14**	**-17.1**	**-2.05**
2	SER112	-6.67	-4.88	-1.80	-7.19	-5.20	-1.20	-8.36	-6.81	-1.55	-9.86	-8.34	-1.53	-8.95	-7.63	-1.32
	SER212	-0.85	-0.78	-0.07	-0.81	-0.78	-0.02	-0.33	-0.28	-0.05	-1.02	-0.45	-0.57	-1.23	-1.22	-0.01
	PRO213	-1.05	-1.05	0	-1.09	-1.09	0	-0.60	-0.58	-0.02	-0.39	-0.36	-0.03	-0.77	-0.75	-0.02
	LEU214	-0.43	-0.35	-0.08	-0.35	-0.32	-0.03	-0.19	-0.14	-0.05	0.03	-0.10	-0.07	-0.10	-0.06	-0.04
	PRO111	-0.37	-0.20	-0.17	-0.24	-0.22	-0.02	-0.50	-0.37	-0.13	-0.50	-0.43	-0.07	-0.70	-0.62	-0.08
	TOTAL	**-9.37**	**-7.26**	**-2.12**	**-9.44**	**-7.61**	**-1.27**	**-9.98**	**-8.18**	**-1.8**	**-11.80**	**-9.68**	**-2.9**	**-11.75**	**-10.28**	**-1.47**

(B)

HYDROPHOBIC POCKET	AMINO ACID	Average Interaction Energy (kcal.mol^{-1})														
		1			2			3			4			5		
		TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC	TOTAL ENERGY	VDW	ELEC
2	LEU272	-0.02	-0.03	0.01	0	0	0	-0.14	-0.13	-0.01	-0.75	-0.71	-0.03	-0.48	-0.43	-0.05
	PRO271	-0.31	-0.29	-0.02	-0.19	-0.21	0.02	-1.08	-1.07	-0.01	-0.98	-0.97	-0.01	-1.13	-1.13	0
	SER270	-0.45	-0.31	-0.14	-0.50	-0.46	-0.04	-1.52	-1.40	-0.12	-0.84	-0.83	-0.01	-1.54	-1.52	-0.02
	HIS268*	-2.03	-0.79	-1.24	3.04	-0.94	-2.10	-1.89	-0.43	-1.47	-1.42	-1.39	-0.03	-1.45	-1.95	0.50
	SER247*	0	0	0	-0.01	-0.01	0	-0.01	-0.01	0	-0.20	-0.20	-0.01	0	0	0
	SER169	-1.55	-1.34	-0.22	-2.87	-1.44	-1.43	-3.10	-1.73	-1.37	-3.79	-2.24	-1.55	-4.33	-2.14	-2.18
	PRO269	-0.92	-0.65	-0.27	-1.35	-1.55	0.20	-2.30	-1.91	-0.39	-1.15	-1.08	-0.07	-2.42	-2.28	-0.14
	TOTAL	**-5.28**	**-3.41**	**-1.9**	**-7.96**	**-4.61**	**-3.79**	**-10.04**	**-6.68**	**-3.36**	**-9.13**	**7.42**	**-1.71**	**-11.35**	**-9.45**	**-1.90**

Table 5.11 Summary of the interaction energy between amino acids which make up hydrophobic pockets 1 and 2 of (A) hsvUDG and (B) hUDG

The smaller volume of the binding pocket and absence of hydrophobic pocket 1 in hUDG are strong reasons as to why the 6-(4-alkylanilino)-uracil inhibitors bind weakly to the hUDG protein. The stable structural presence of hydrophobic pocket 1 in hsvUDG, located adjacent to the binding pocket, provides ideal hydrophobic interactions for the alkyl chain of 6-(4-alkylanilino)-uracil inhibitors of sufficient length. Table 5.11 illustrates the interaction energy between the inhibitors and specific amino acids which form hydrophobic pocket 1 and 2 in both hUDG and hsvUDG. The interaction energies determined for the hydrophobic pockets are in agreement with the FEP results. Despite the evolutionary similarities in the structure of the hUDG and hsvUDG proteins, the inhibitors are able to selectively inhibit the hsvUDG protein and have very little effect on the hUDG protein.

References:

1. T. Lindahl, *Proc. Natl. Acad. Sci U. S. A.*, 1974, **71**, 3649-3653.
2. A. Verri, P. Mazzarello, G. Biamonti, S. Spadari and F. Focher, *Nucleic Acids Res.*, 1990, **18**, 5775-5780.
3. R. B. Pyles and R. L. Thompson, *Journal of Virology*, 1994, **68**, 4963-4972.
4. R. Chen, H. Wang and L. M. Mansky, *Journal of General Virology*, 2002, **83**, 2339-2345.
5. K. Krusong, E. P. Carpenter, B. S. R. W., R. Savva and G. S. Baldwin, *Journal of Biological Chemistry*, 2006, **281**, 4983-4992.
6. H. M. Sun, C. X. Zhi, G. E. Wright, D. Ubiali, M. Pregnolato, A. Verri, F. Focher and S. Spadari, *Journal of Medicinal Chemistry*, 1999, **42**, 2344-2350.
7. O. M. Becker, A. D. MacKerell, B. Roux and M. Watanabe, *Computational Biochemistry and Biophysics*, Marcel Dekker, New York, 2001.
8. B. R. Brooks, R. E. Bruccoleri, O. B. D., D. J. States, S. Swaminathan and M. Karplus, *Journal of Computational Chemistry*, 1987, **4**, 187-217.
9. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *Journal of Chemical Physics*, 1983, **79**, 926-935.
10. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez and J. A. Pople, Gaussian, Inc., Wallingford CT, 6th edn., 2004.
11. N. Foloppe and A. D. MacKerel, *Journal of Computational Chemistry*, 2000, **21**, 86-104.
12. C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, 2nd edn., John Wiley and Sons Inc., 2004.
13. A. D. MacKerell and N. K. Banavali, *Journal of Computational Chemistry*, 2000, **21**, 105-120.
14. A. R. Leach, *Molecular Modelling: Principles and Applications*, Prentice Hall, New York, 2001.
15. W. D. Rogers, *Computational Chemistry Using the PC*, 3rd edn., John Wiley and Sons Inc., New Jersey, 2003.
16. C. Chipot and A. Pohorille, *Free Energy Calculations: Theory and Applications in Chemistry and Biology*, Springer, New York, 2007.
17. S. S. Parikh, C. D. Mol, G. Slupphaug, S. Bharati, H. E. Krokan and J. A. Tainer, *Embo Journal*, 1998, **17**, 5214-5226.

18. R. Savva, K. McAuleyhecht, T. Brown and L. Pearl, *Nature*, 1995, **373**, 487-493.
19. J. Kuszewski, A. M. Gronenborn and M. G. Clore, *Protein Science*, 2008, **5**, 1067-1080.
20. J. E. Nielsen and G. Vriend, *Proteins*, 2001, **43**, 403-412.
21. Glide, version 5.6, Schrodinger, Inc., New York, 2010.

Chapter 6

Conclusions and Future Work

Although alignment analyses revealed that the enzymes are 40.1% identical and the binding pockets are highly conserved, binding structure analyses shows a very different behaviour towards the DNA substrate. From DNA substrate simulations, it can be seen that the sugar-phosphate backbone plays an important role in the binding of the substrate to the UDG enzymes. The interactions between the DNA substrate and the UDG enzymes are comprised primarily of strong non-specific electrostatic and van der Waals interactions. The human uracil-DNA glycosylase enzyme (hUDG) forms a stronger interaction with the DNA substrate than the herpes simplex virus type 1 uracil-DNA glycosylase enzyme (hsvUDG). The volume of the hUDG enzyme binding pocket was determined to be approximately 20Å smaller than the hsvUDG enzyme. Based on the distance between the uracil head and the ASN204 and ASN147 amino acids (which are considered the base of the binding pocket), it can be seen that the uracil in hUDG enters the binding pocket deeper than in hsvUDG. The DNA substrate simulations provide insight into the natural behaviour of the enzymes.

The 6-(4-alkylanilino)-uracil inhibitors selectively inhibit the herpes simplex virus type 1 uracil-DNA glycosylase enzyme (hsvUDG). From this study it can be concluded that the presence of the conformationally stable hydrophobic pocket 1 in the hsvUDG enzyme is a fundamental reason for this selectivity. The length of the alkyl chain in the 6-(4-alkylanilino)-uracil inhibitors has to be of the correct length in order to ensure optimal binding in hydrophobic pocket 1. The 6-(4-octylanilino)-uracil inhibitor, or inhibitor 4, seems to be of optimal length to bind strongly with the hsvUDG enzyme. Alkyl chains in the simulations of the inhibitors in the hUDG enzyme displayed a preference for hydrophobic pocket 2. Interaction energies between the inhibitor and the hsvUDG enzyme are on average 12.6kcal.mol^{-1} greater than the interaction energies of the inhibitors in hUDG. The smaller volume of the binding pocket and absence of a stable hydrophobic pocket 1 in

Appendix

A.1 Root Mean Square Deviation (RMSD) Plots

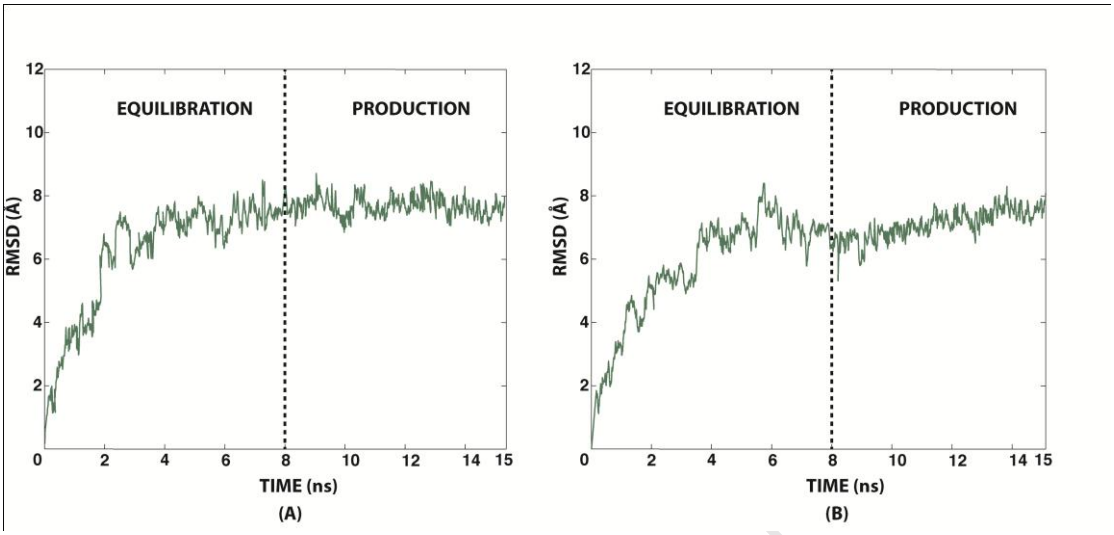


Figure A.1 RMSD for DNA substrate simulations in the binding pocket of (A) human uracil-DNA glycosylase and (B) herpes simplex virus type 1 uracil-DNA glycosylase.

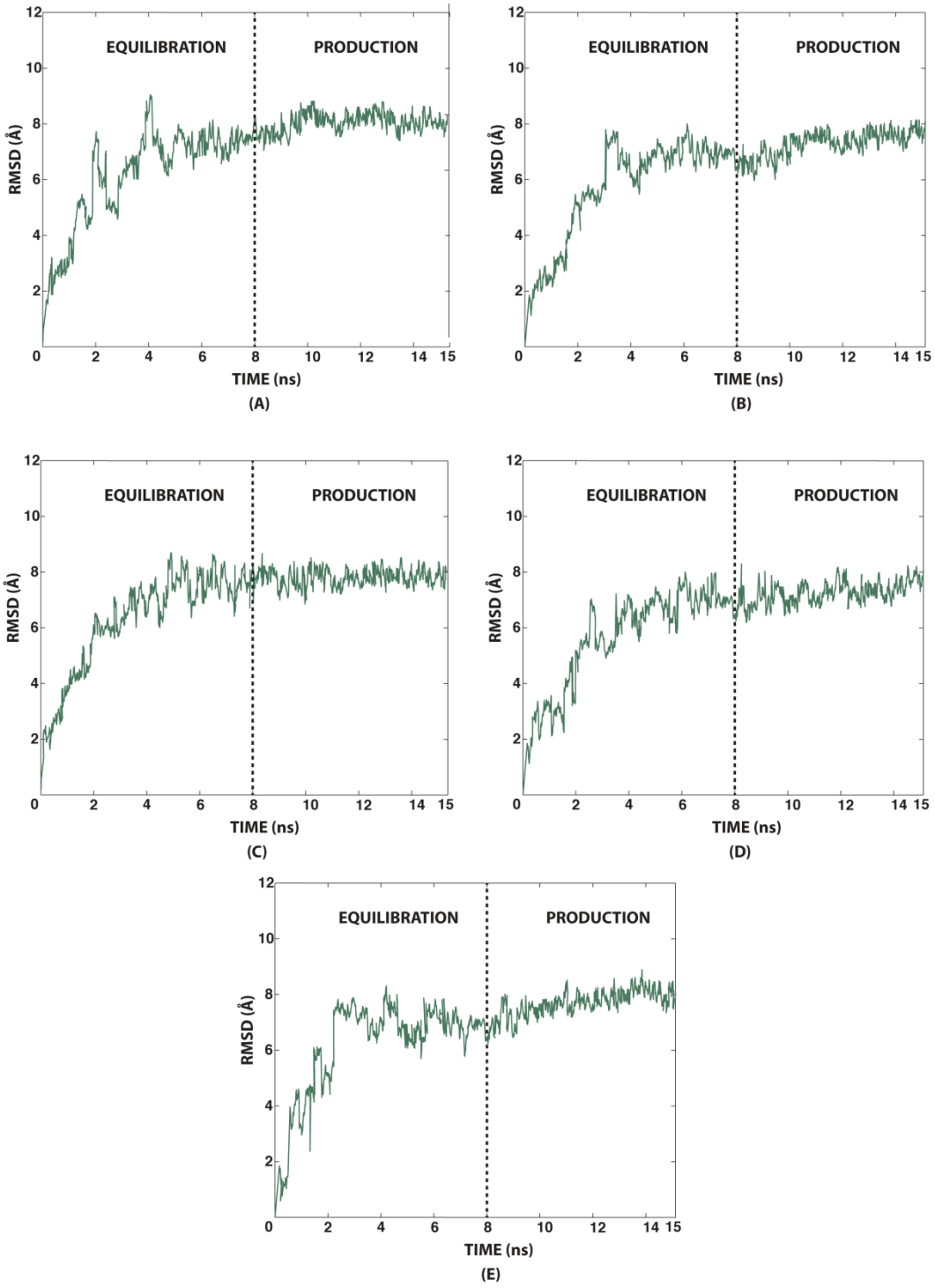


Figure A.2 RMSD for human uracil-DNA glycosylase enzyme simulations with (A) inhibitor 1 (B) inhibitor 2 (C) inhibitor 3 (D) inhibitor 4 (E) inhibitor 5, as substrates in the binding pocket of the enzyme.

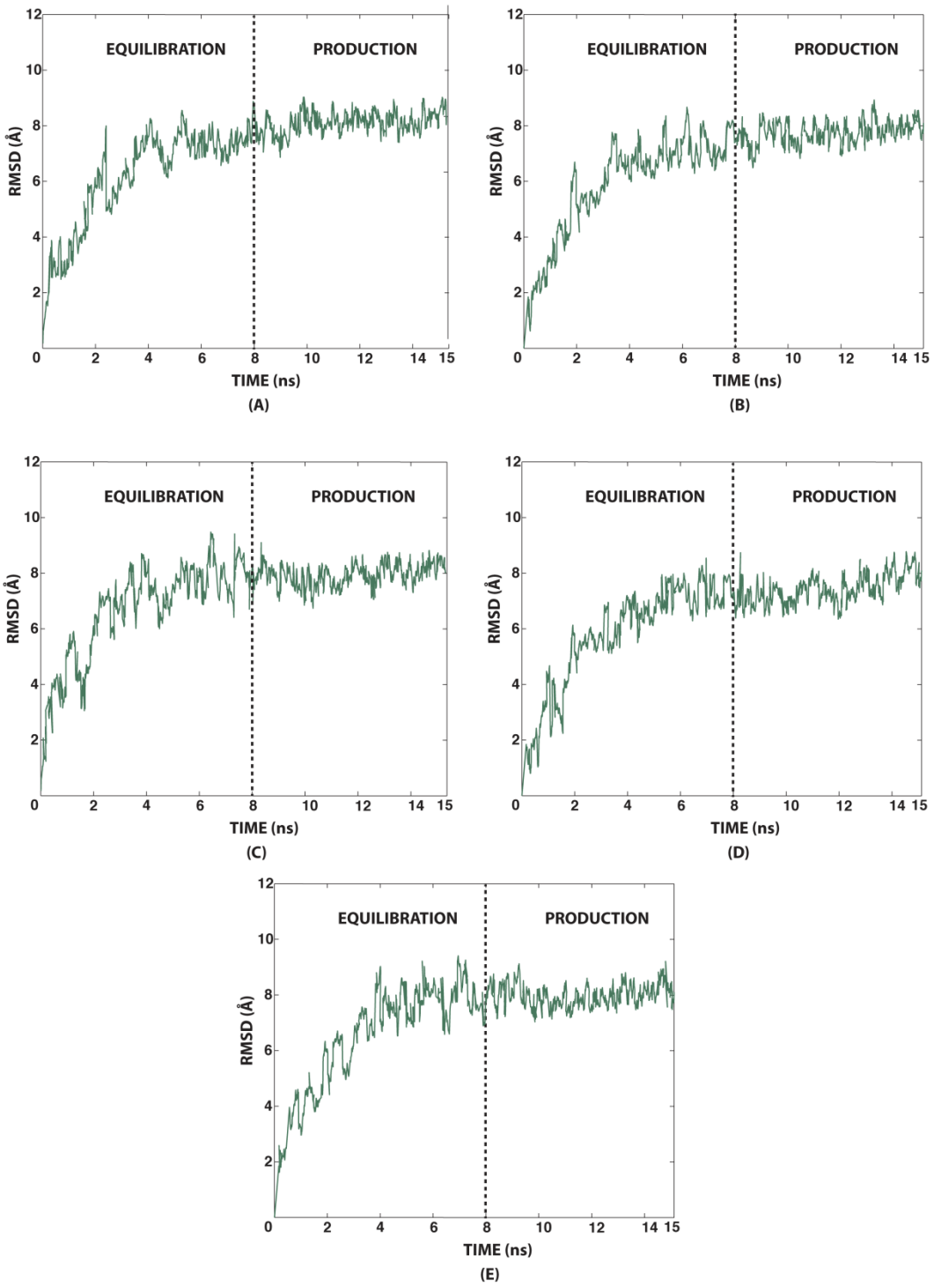


Figure A.3 RMSD for herpes simplex virus type 1 uracil-DNA glycosylase enzyme simulations with (A) inhibitor 1 (B) inhibitor 2 (C) inhibitor 3 (D) inhibitor 4 (E) inhibitor 5, as substrates in the binding pocket of the enzyme.

hUDG are strong reasons as to why the 6-(4-alkylanilino)-uracil inhibitors bind weakly to the hUDG protein. RMSD measurements indicate that the alkyl chain moves more freely when bound in hydrophobic pocket 2 than in hydrophobic pocket 1.

Using the binding model determined in this study, improved inhibitors can be developed and physicochemical properties of the inhibitors can be improved upon. Further quantum mechanical studies can be carried out on these inhibitors to gain further insight into their behaviour.