RATIONALISING THE EFFECT OF GUANINE-BASED RICIN INHIBITORS USING COMPUTER SIMULATIONS

RANGA S. JAYAKODY

RATIONALISING THE EFFECT OF GUANINE-BASED RICIN INHIBITORS USING COMPUTER SIMULATIONS

Dissertation presented to the

UNIVERSITY OF CAPE TOWN

In fulfilment of the requirements for the degree of

MASTER OF SCIENCE

By

RANGA S. JAYAKODY
BSc Hons. (Carleton University, Canada)

Supervisor: Assoc. Professor Kevin J. Naidoo

Department of Chemistry
University of Cape Town
2008

DECLARATION

I declare that RATIONALISING THE EFFECT OF GUANINE-BASED RICIN INHIBITORS USING COMPUTER SIMULATIONS is my own work and that all the sources that I have used or quoted have been indicated and acknowledged my means of complete references.

Signed by candidate

Rānga Ś. Jayakody

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Associate Professor Kevin J. Naidoo for invaluable academic guidance, motivation, funding and great understanding.

Dr. Shirley Churms greatly assisted me in improving my writing skills by proof reading and making valuable suggestions. Thank you very much.

This work would never have been possible without the endless support of my parents throughout my academic career. Thank you so much.

A special thank is required for my wife Anusha for her patience, enthusiastic encouragement and invaluable support in all aspects. A unique thank goes to my little son Yasith for letting dad out of home for most of the day. I love you both very much.

My dear colleague Richard Matthews deserves special thanks for proof reading and his valuable suggestions. I would also like thank my other group members, Dr. Gerhard Venter and Chris Barnett for their help and support.

Abstract

Ricin is a plant toxin isolated from the castor bean plant. It has been ranked among the most toxic substances known with a range from 0.1-1.0 μg / kg for humans. This has lead to its use as a biological weapon and a therapeutic agent. The reaction mechanism of ricin is not known and so rational design is limited, resulting in no effective inhibitor being found for this enzyme. An important step towards understanding the reaction mechanism and the design of effective inhibitors is the complete understanding of the binding pocket and the mode / nature of binding of ligands to the protein.

In this thesis computational methods were used to rationalise the effect of selected ricin inhibitors. Free Energy Perturbation (FEP) methods (described in Chapter 3) were used to determine the relative binding free energies of five guanine-based inhibitors. The free energy results provide a ranking that is in the same order as that found from the relative IC_{50} values. The new force field parameter set developed here (Chapter 4) and introduced into the exsisting CHARMM27 force field was consequently validated.

These parameters were used to produce 15 ns long Molecular Dynamics simulations (MD) for inhibitor:ricin and substrate (adenosine):ricin complexes. The analyses of the trajectories from MD simulations (Charter 4) allowed us to identify key amino acids in the ricin binding site. It was found that the mode of binding to the substrate and the best inhibitors was primarily via electrostatic interactions. Two regions of interaction were identified in the binding site. While the substrate interacted with both regions the inhibitors studied here interacted with only one of those regions. The region to which the inhibitors bind has an overall positive electrostatic potential, which compliments their overall negative potential leading to relatively large inhibitory activity. This study reports the key amino acids in the binding pocket and rationalises the relative inhibition of ricin resulting from guanine-based ligands.

ABBREVIATIONS

2D	Two Dimensional
3D	Three Dimensional
Å	Angstroms
BER	Base Excision Repair
CADD	Computer Aided Drug Design
CHARMM	Chemistry at Harvard Macromolecular Mechanics
IC_{50}	Median Inhibition Concentration
COG	Centre of Geometry
DFT	Density Functional Theory
DNA	Deoxyribonucleic Acid
EF	Elongation Factor
EM	Energy Minimisation
EPS	Electrostatic Potential Surface
FEP	Free Energy Perturbation
LD	Langevin Dynamics
MC	Monte Carlo
MD	Molecular Dynamics
MEP	Molecular Electrostatic Potential
mM	Milimoles
MM	Molecular Mechanics
m-RNA	Messenger Ribonucleic acid
NMR	Nuclear Magnetic Resonance
ns	Nanosecond
NVT	Canonical ensemble
PMF	Potential of Mean Force
ps	Picosecond
QM	Quantum Mechanics
RIP	Ribosome Inactivating Protein
RMSD	Root Mean Square Deviation
RNA	Ribonucleic Acid
r-RNA	Ribosomal Ribonucleic Acid
RTA	Ricin A chain
SBDD	Structure Based Drug Design
SE	Semi empirical
SG	Slow Growth
TI	Thermodynamic Integration
vdW	van der Waals

CHAPTER 1

Enzymes

1.1 Introduction to Enzymes

Chemical reactions play crucial roles in living organisms. The collection of biochemical reactions essential to maintaining life known as metabolism[1] occurs at the cellular level and it is a continuous process throughout the life of an organism. Some metabolic processes involve the build-up of new tissues, replacement of old tissues, conversion of food to energy, disposal of waste materials, and reproduction[1].

The majority of biochemical reactions are non-spontaneous making the system more viable by preventing misplaced and untimely reactions from occurring. Nature employs the phenomenon of catalysis to make these reactions feasible at the appropriate locations and times. A catalyst increases the rate of a chemical reaction without itself undergoing any permanent chemical change. Catalysts of biochemical reactions are referred to as enzymes, which are a special class of proteins. In the absence of enzymes the biochemical reactions take place at extremely low rates which are far too low for the pace of metabolism[2].

Figure 1.1 Lock and Key mechanism of enzymatic catalysis.

Enzymes are very efficient catalysts, often far superior to conventional chemical catalysts, and therefore are increasingly used in a wide range of industries. Some of the other significant differences between conventional chemical catalysts and enzymes are their regiospecificity, stereospecificity and substrate specificity.

Because of this specificity, a chosen reaction is catalysed by enzymes to the exclusion of side-reactions, eliminating undesirable by-products. The specificities of enzymes are generally explained in terms of to the "lock and key" mechanism. A schematic representation of this mechanism is given in Figure 1.1.

1.2 Structure of Enzymes

The major constituent of enzymes is protein; generally enzymes are globular proteins. The primary structure of an enzyme is defined as the amino acid sequence of its polypeptide chain or chains. The higher levels of protein structure – secondary, tertiary and quaternary – refer to the three-dimensional shapes of folded polypeptide chains. The activities of enzymes are governed by their three-dimensional structures.

The catalytic activity is performed by only a small number of amino acids in an enzyme. The region of the structure that contains the catalytic amino acids is referred to as the active site. In the active site the functional groups are arranged in such a way that they can take part in a specific chemical reaction. Some enzymes are composed purely of proteins while others contain some non-protein moieties, which are usually either carbohydrates, metal ions (Fe^{3+} ,Zn^{2+} , Cu^{2+}), co-enzymes or a combination of these species, which may or may not participate in the catalytic activity of the enzyme[2]. Generally, covalently attached carbohydrate groups that are found in proteins have no direct influence on the catalytic activity, but they are related to the enzyme's stability and solubility. The metal ions that are bound to an enzyme via noncovalent or covalent bonds are called cofactors and they are often important to both the activity and the stability of the enzymes.

1.3 The Mechanism of Enzyme Catalysis

1.3.1 Energetics

In any chemical reaction, the reactant molecules must contain sufficient energy to cross a potential energy barrier, i.e. the activation energy, in order for the reaction to occur. In general, the intrinsic energy of most molecules is not sufficient to overcome the activation energy.

It is therefore necessary either to lower the potential energy barrier to the reaction or to supply more energy to the reactants so that the reaction may be propelled over the barrier.

The enzyme energetics are best explained by the Transition State (TS) Theory[3, 4]. In the TS theory, the only physical entities considered are the ground state and the transition state (the most unstable species on the reaction pathway). The TS occurs at the peak of the energy profile of an enzymatic reaction, implying its high instability. In the transition state the chemical bonds are in the process of being made and broken. The energetics of a typical enzymatic reaction is illustrated in Figure 1.2 where E and **S** are the enzyme and the substrate (reactants) respectively. In the absence of the enzyme, the reaction proceeds via transition state **TSu** with activation energy of G_u^* (dotted line). The rate of the reaction shown in Figure 1.2 is given by equation 1.1 where k is the ordinary rate constant of the elementary reaction and k' the rate constant for the decomposition of the TS_u to products.

$$\frac{d[P]}{dt} = k[E][S] = k'[TS_u] \qquad \textbf{(1.1)}$$

In TS theory, it is assumed that the transition state is in thermodynamic equilibrium with the ground state, that is,

$$K^* = \frac{[TS_u]}{[E][S]} \qquad \textbf{(1.2)}$$

where K^* is an equilibrium constant, which can be expressed as

$$-RT \ln K^* = \Delta G_u^* \qquad \textbf{(1.3)}$$

where T is the absolute temperature and R is the gas constant .
The combination of the equations 1.1 to 1.3 yields equation 1.4, where k' is defined as $k' = k_B T/h$; in which k_B, T and h are Boltzmann constant , absolute temperature and Planck's constant respectively.

$$\frac{d[P]}{dt} = k' e^{-\Delta G_u^*/RT}[E][S] \qquad \textbf{(1.4)}$$

The equation 1.4 indicates that the rate of the reaction depends not only on the concentration of its reactants, but also decreases exponentially with G_u^*. Thus, the larger the free energy difference between the transition state and the reactants, that is, the less stable the transition state, the more slowly the reaction proceeds. Enzymes function by forming a transition state with the reactants of lower free energy than would be found in the uncatalysed reaction and so increase the reaction rate. This is referred to as transition state stabilisation mechanism. The principle TS stabilisation mechanisms used by enzymes are proton transfer and electrostatic stabilisation.

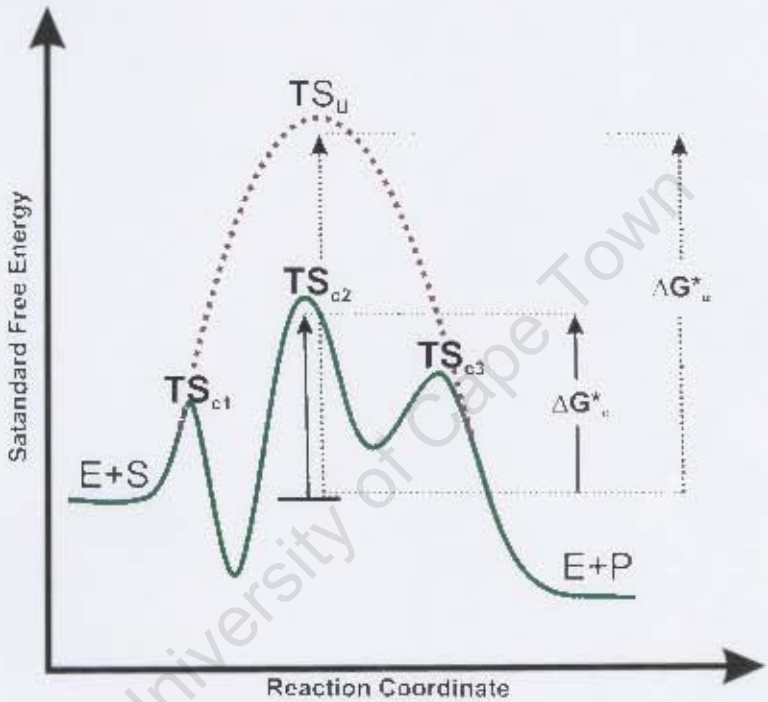


Figure 1.2 Free energy profile of an enzyme-catalysed reaction

The activation energy barrier of an enzymatic reaction can also be lowered by means other than TS stabilisation. In this context, the most important mechanism is the initial binding of substrate to its enzyme in an orientation that will facilitate the reaction. The binding energy released from this process is partially utilised to reduce the opposing activation entropy barrier. The activation entropy originates from the loss of the reactant and catalytic groups' translational and rotational entropy. Other factors contributing to the lowering of activation entropy are the introduction of strain into the reactants (allowing more binding energy to be available for the transition state) and the desolvation of reacting and catalysing ionic groups.

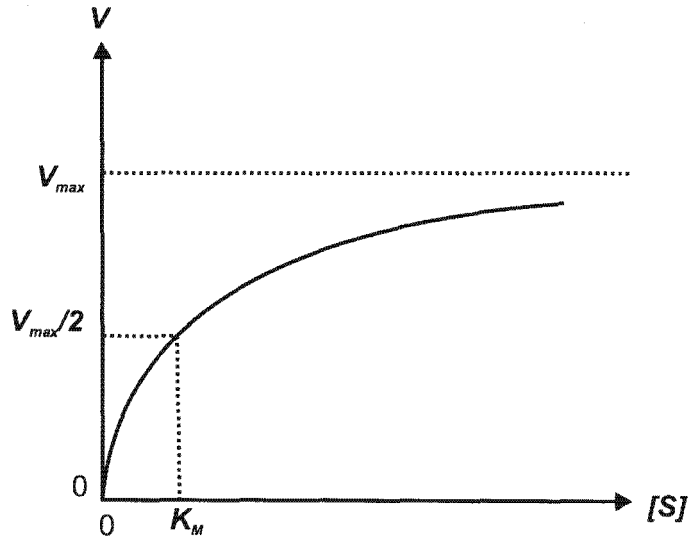


Figure1.3 A Plot of the velocity (v) of a simple enzymatic reaction versus substrate concentration

The Michaelis constant K_M in equation 1.5 is defined as the substrate concentration at which the reaction velocity is half V_{max}. The kinetic phenomena for a single substrate enzymatic reaction can be explained by the Michaelis-Menten mechanism (Figure 1.4)

$$E + S \underset{}{\overset{K_S}{\rightleftharpoons}} ES \xrightarrow{k_{cat}} E + P$$

Figure 1.4 A Typical enzymatic reaction

where the catalytic reaction is divided into two processes. First the enzyme (E) and substrate (S) form an enzyme-substrate complex (ES) and this step is assumed to be rapid, reversible and chemically inactive, i.e. no chemical reaction takes place in this step. The ES is then converted into product by means of chemical reactions with a first order rate constant k_{cat}. The equilibrium constant for the first stem is given by K_S.

From Figure 1.4 ,

$$K_S = \frac{[E][S]}{[ES]} \qquad (1.6)$$

and

$$Rate(v) = k_{cat}[ES] \qquad (1.7)$$

Also, the total enzyme concentration $[E]_T$ and the concentration of the free enzyme [E] are related by

$$[E] = [E]_T - [ES] \qquad (1.8)$$

when V_{max} is given as

$$V_{max} = k_{cat}[E]_T \qquad (1.9)$$

From the equations 1.6 to 1.9, it is possible to show that:

$$v = \frac{V_{max}[S]}{K_S + [S]} \qquad (1.10)$$

Equation 1.10 is identical to equation 1.5 where K_M is equal to the disassociation constant of the enzyme-substrate complex K_S. The enzyme-substrate complex is the foundation of enzyme kinetics and is often termed the Michaelis complex. The kinetics of all enzymatic reactions are studied with the Michaelis-Menten mechanism or with extended and / or modified Michaelis-Menten mechanisms such as the Briggs-Maldane and Briggs-Haldane mechanisms[4].

1.3.3 Inhibition

The activity of enzymes can be reduced by certain substances, which make reversible complexes with enzymes, thus influencing the binding of the substrate. Substances that reduce enzyme activity in this fashion are termed inhibitors. There are four main types of inhibition; competitive, non-competitive, uncompetitive and mixed inhibition.

(a) Competitive Inhibition

If an inhibitor (I) binds to the active site of the enzyme reversibly and prevents the binding of the substrate (S), I and S are competing for the active site and I is termed a competitive inhibitor.

Figure 1.5 Michaelis-Menten mechanism of competitive inhibition

In competitive inhibition, a simple Michaelis-Menten mechanism (Figure 1.4) will have an additional equilibrium which is represented by the equilibrium constant K_I in Figure 1.5. The overall rate of a competitive inhibition reaction is given as (based on Michaelis-Menten equation)

$$v = \frac{[E]_0[S]k_{cat}}{K_M(1+[I]/K_I)+[S]} \qquad (1.11)$$

(b) Non-competitive, Uncompetitive and Mixed Inhibition

When an inhibitor (I) and the substrate (S) simultaneously bind to the enzyme instead of competing for the same active site, different inhibition patterns occur (Figure 1.6).

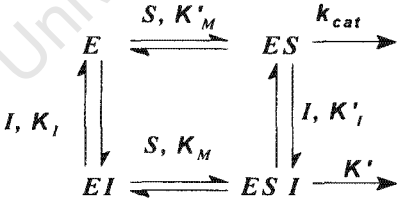


Figure 1.6 Non-competitive, Uncompetitive and Mixed Inhibition

In non-competitive inhibition, *ESI* does not react (i.e. $K'=0$) and the dissociation constants of S from *ES* and *ESI* are the same (i.e. $K_M = K'_M$). An inhibition in which both k_{cat} and K' are affected is termed mixed inhibition. Uncompetitive inhibition occurs when an inhibitor binds to ES but not to E.

1.4 Classes of Enzymes

The enzymes are classified into six main classes according to the reactions that they catalysed. These classes are given below.

I. **Oxidoreductases:** This class of enzymes catalyses redox reactions where there is a transfer of hydrogens, oxygens or electrons between molecules. This extensive class includes the dehydrogenases (hydride transfer), oxidases (electron transfer to molecular oxygen), oxygenases (oxygen transfer from molecular oxygen) and peroxidases (electron transfer to peroxide).

II. **Transferases** : The reactions where transfer of an atom or group of atoms (e.g. acyl-, alkyl- and glycosyl-), between two molecules (excluding such transfers as are classified in the other groups) takes place are catalysed by this class of enzymes.

III. **Lyases:** Elimination reactions in which a group of atoms is removed from the substrate are catalysed by lyases. This class of enzymes includes the subclasses aldolases, decarboxylases, dehydratases and some pectinases.

IV. **Isomerases:** This class of enzymes catalyses molecular isomerisation reactions. This includes epimerases, racemases and intramolecular transferases.

V. **Ligases:** This category of enzymes is also known as synthetases. The biochemical reactions that involve the formation of a covalent bond joining two molecules together are catalysed by these enzymes.

VI. **Hydrolases:** This class of enzymes catalyses the reactions which involve hydrolytic reactions and their reversal. This is the most commonly encountered class of enzymes in biological systems and therefore has a vast range of catalytic activity and a great deal of importance. Based on the bonds that they act upon, hydrolases can be further classified into several subclasses, which are given in Table 1.1.

Among the many types of hydrolase enzymes, glycosylases are of prime importance as they are involved in many biochemical reactions which are crucial to the existence of organisms. Special attention will be given to glycosylases later in this chapter as the enzyme studied for this thesis belongs to that category.

Subclass	Type of bond acted on
1	bonds in sugars (glycosylases)
2	carbon-nitrogen bonds, other than peptide bonds
3	carbon-carbon bonds
4	phosphorus-nitrogen bonds
5	carbon-phosphorus bonds
6	carbon-sulfur bonds
7	ester bonds (esterase)
8	peptide bonds (peptidases)
9	acid anhydrides
10	halide bonds
11	sulfur-nitrogen bonds
12	sulfur-sulfur bonds

Table 1.1 Subclasses of Hydrolases

1.5 Glycosylases

Glycosylases are a subclass of hydrolases and can be categorised further according to the type of glycosidic bond that they cleave. These categories are glycosidases, the enzymes hydrolysing *O*- and *S*-glycosyl compounds and *N*-glycosidases, the enzymes hydrolysing *N*-glycosyl compounds.

A glycosidic bond is formed when the anomeric group of a sugar condenses with an alcohol to form alpha and beta glycosides which are cyclic acetals or ketals. An *N*-glycosidic bond is formed when the anomeric carbon forms a bond with an amine, and they are the bonds that connect D-ribose to purines and pyrimidines in the nucleic acids. These two types of bonds are shown in Figure 1.7 Under normal physiological conditions, the hydrolysis of these bonds is extremely slow, necessitating the action of glycosylases. The two subclasses of glycosylases play crucial roles at two distinct locations in metabolism. Glycosidases are involved in detaching required monosaccharides from their polysaccharides and N-glycosidases are mainly involved in manipulating genetic material (DNA / RNA).

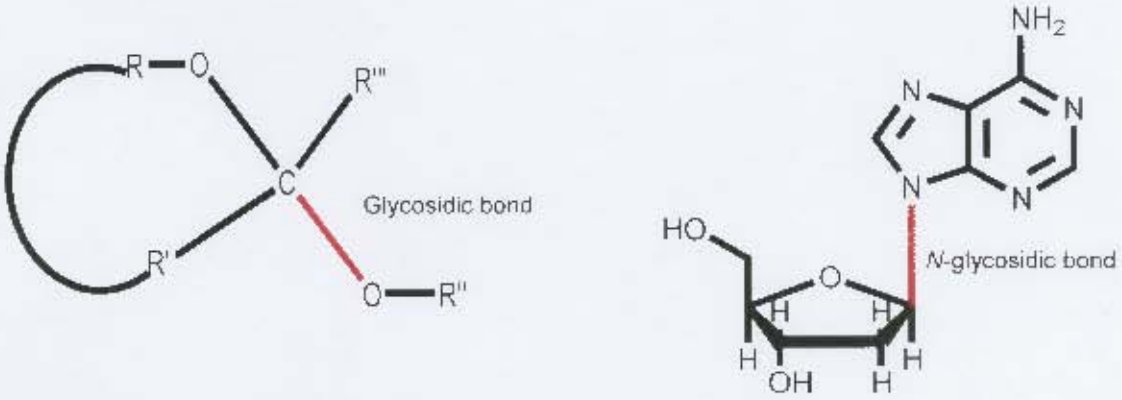


Figure 1.7 Different types of glycosidic bonds

1.5.1 N-Glycosidases and Base Excision

Damage to genetic material in a living cell can occur as a result of various endogenous and exogenous processes. Under normal conditions endogenous processes are the major source of damage [5] and these processes include oxidation by peroxides, spontaneous damage , methylation and so forth. The resulting lesions dramatically decelerate or prevent the cellular replication and some lesions create mutations, thereby generating a cytotoxic effect.

Repairing a damaged base in nucleic material (DNA and RNA) is critical to cellular replication. A combination of enzymes is responsible for the base repair mechanism in the cell. With most types of damage, the base repair mechanism is initiated by N-glycosylases. These enzymes catalyse the hydrolysis of the N-glycoside bond between a damaged base and the sugar-phosphate backbone. The type of base repair initiated by glycosylases is known as base excision repair (BER), because it removes free bases from nucleic material. Thus BER mechanism protects the cell from cytotoxic effects exerted by damaged bases.

Although base excision is a mechanism essential to cell function, it can be fatal if an undamaged base of the original genetic sequence is removed. Ribosome Inactivating Proteins (RIP) are a class of enzyme (glycosylases) that operates in the above fashion, and these are well known as cytotoxins. This thesis is focused on a member of RIP family and RIPs are discussed below.

1.6 Ribosome-Inactivating Proteins (RIP)

Ribosome-inactivating proteins (RIPs) are a special group of base excision enzymes (*N*-glycosidases) that distinctively cleave nucleo-bases and render them inactive. RIPs are therefore referred to as cytotoxins. This class of proteins is generally found in plants. Some of many plants[6] that produce RIP and accumulate them are; *Ricinus communis* (castor bean plant), *Abrus precatorius* (jequirity bean plant), *Trichosanthes kirilowii* and *Momordica charantia*. The RIPs have a wide distribution in nature. They are present mostly in Angiospermae, both mono- and dicotyledons[6], in mushrooms[7] and in an alga, *Laminaria japonica*[8].The amount of RIPs in plant tissues is highly variable, ranging from traces to hundreds of milligrams per 100 g[6]. The RIPs can be found localised to one type of tissue or distributed in different types of tissues.

Ribosome inactivating proteins are classified into three types. Type I is composed of a single polypeptide chain, whereas type II is a heterodimer consisting of an A chain, functionally equivalent to a type I, which is attached to a sugar-binding B chain. Type III is a single chain containing an extended carboxyl-terminal domain with unknown function. A schematic representation of these different types of RIPs is given in Figure 1.8 (adopted from [6]). Although, these plant proteins are known for their activity of depurinating ribosomes at the sarcin/ricin (S/R) loop of the r-RNA, there is no clear answer to the question of why plants synthesise and accumulate RIPs. Interestingly, their biological action still remains open to speculation although RIPs were first identified more than 100 years ago.

1.7 Ricin

Ricin which is one of the best known RIPs today, is isolated from *Ricinus communis* (Castor bean plant) [9, 10]. It was first identified and named "Ricin" after the work of H. Stillmark in 1888 [11]. As this was the first time a well-defined biological activity had been assigned to a plant protein, the identification of Ricin was an important milestone in biochemistry.

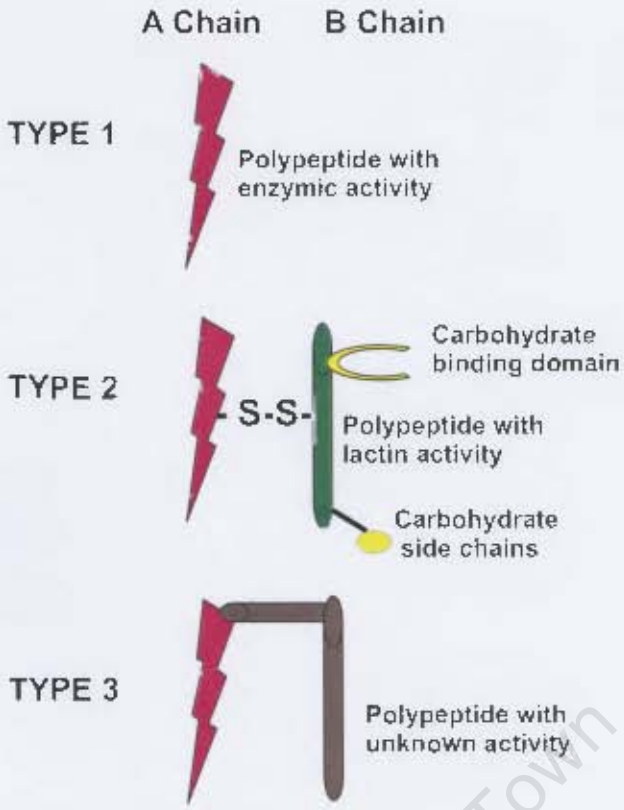


Figure 1.8 Types of Ribosome Inactivating Proteins.

The current applications of ricin range from therapeutic agents such as immunotoxins[12] to biological weapons. The castor bean plant, castor seeds and the three dimensional structure of ricin are presented in Figure 1.9. The toxic dose of ricin for humans is likely to be in the 0.1-1.0 µg/kg range, and it has been ranked among the most toxic substances known[10].

(a) (b) (c)

Figure 1.9 (a). Castor Bean Plant **(b).** Castor Seeds **(c).** Ricin-space filling structure

Ricin is a heterodimeric (A-B type) protein and its capability of blocking the mechanism of protein synthesis was first identified by Lin et al. in 1977[13]. The ricin A-chain (RTA) shows the catalytic activity while the B-chain facilitates the direction of the A-chain to the cell surface. The structure of RTA is given in Figure 1.10. The A chain is released to the cell by cleaving the disulfide bond and it is then internalised by endocytosis[14]. The RTA is then directed to the endoplasmic reticulum via retrograde transport[15] and subsequently it is moved to the cytosol wherein it binds to a specific nucleotide sequence on the sarcin-ricin tetra loop of the 28S ribosomal RNA.

Figure 1.10 Ricin A chain

The GAGA tetra loop of the stem-loop structure of ribosome is specifically depurinated by RTA catalysis[16]. A^{4324} is the specific adenine of the 28S ribosomal RNA (rat) that is removed by RTA catalysis[9]. The r-RNA target site for ricin is schematically presented in Figure 1.11.

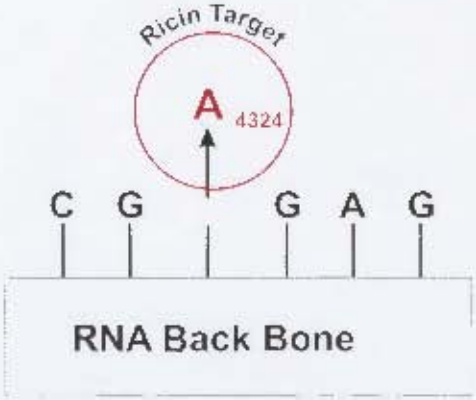


Figure 1.11 r-RNA target for Ricin

Removal of adenine 4324 destroys an elongation factor (EF) binding site [17] of the ribosomal RNA (Figure 1.12 (a)). As a result, the movement of m-RNA / ribosome is lost. Consequently, the protein synthesis ceases and cell death is caused. A schematic representation of the arresting of the protein synthesis by Ricin is given in Figure 1.12.

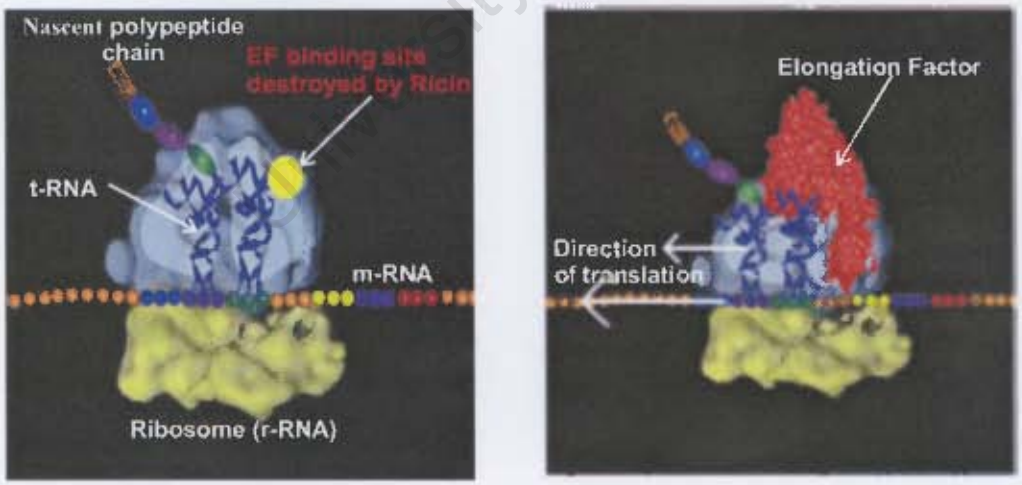


Figure 1.12 Ceasing Protein Synthesis by Ricin.

Binding of elongation factor does not occur as shown in figure (b) when ricin destroys the EF site shown in (a). The loss of mobility of m-RNA / ribosome causes the whole protein synthesis mechanism to cease.

Identification of its target adenine by ricin among millions of other nucleo bases is quite intriguing and the exact process is yet unknown. It has been suggested that ricin might use a base flipping mechanism to dock its target adenine to the active site[18]. According to the proposed mechanism, RTA catalytic reaction proceeds via an S_N1 mechanism, forming an oxacarbenium intermediate followed by addition of water in a separate step [19, 20]. The formation of the transition state at the active site involves leaving group activation, oxacarbenium ion formation and generation of the incipient water nucleophile[21]. It is assumed that the leaving adenine is partially protonated by ARG 180. Residue GLU 177 is assumed to be stabilizing the oxacarbenium ion and to act as a base to polarise the attacking water[22]. The complete reaction mechanism of ricin still remains unknown. The proposed mechanism is presented in Figure 1.13.

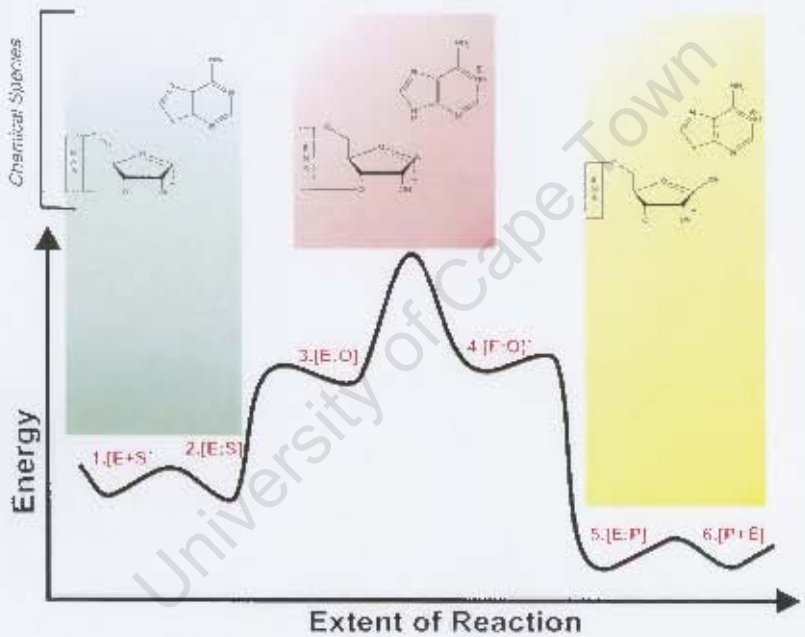


Figure 1.13 Proposed Mechanism for RTA catalysis.

Free Ricin A chain (RTA , E in the figure) and RNA substrate (S)- (1). Michaelis complex- (2) Noncovalent RTA : (oxocarbenium ion + adenine) complex- (3).Unknown intermediate complex (Structure depends on the non chemical step- (4). RTA: (depurinated RNA + adenine) complex-(5) .Free RTA and products- (6).

The crystal structure of ricin has been solved[23] refined[24] and described[24]. From the x-ray studies, several key amino acid residues have been identified; these include GLU177 , ARG180, TRP211, TYR80 and TYR123.

It is proposed [25] that TYR80 and TYR123 sandwich the base in between their phenyl rings. Several other kinetic studies and direct mutagenesis experiments have proposed activities for these residues in the catalytic mechanism[10]. ARG180 is expected to play a role in ground state stabilisation (electrostatic attraction) of ribose and / or transition state destabilisation (repulsion) of the oxacarbenium ion[21]. GLU177 is expected to stabilise the transition state or to act as the base in the catalytic mechanism[22]. The role of TRP211 is non specific. GLU 208 is another residue that has been implicated in catalysis[26]. However, understanding about the binding pocket of ricin and the role of individual amino acids in catalysis still remains incomplete.

1.8 Objectives

The principle objective of this thesis is to rationalise the inhibitory effect of ricin guanine-like inhibitors. The complete understanding of these inhibitors can be used in proposing the essential features of better ricin inhibitors. In order to understand how the enzyme works on different ligands, it is mandatory to know the structure of the enzyme entirely. Additionally, it also requires the understanding of the structures of the complexes of the enzyme with its reactants, intermediates, inhibitors and the products. When this information is available, it is possible to see what catalytic groups are closer to a ligand and what structural changes occur in the ligand and the binding site on binding.

To reveal the above information about ricin this thesis employs two distinct computational techniques; Free Energy Perturbation (FEP) and Molecular Dynamics (MD) simulations. The FEP methods will be used to calculate the relative free energy of binding for selected guanine-like inhibitors which makes possible ranking of the inhibitors (according to their inhibition power). The methods that are used in FEP calculations are presented in chapter 3. The structural and the energetic information of the ricin: ligand (substrate, inhibitors and products) complexes will be explored using MD simulation methods. Molecular Dynamics simulation methods are discussed in chapter 2. The results are discussed in chapter 4 and chapter 5 discusses the future scope of the topic.

CHAPTER 2

Simulation of Proteins

2.1 Introduction

Studying the structure and the function of biomolecules is a major challenge in modern-day science. Among many important biomolecules, enzymes receive special attention owing to their crucial roles in regulating the biochemical reactions. Experimental techniques such as x-ray crystallography, fluorescence [27, 28], infrared spectroscopy [29, 30], NMR[31] and kinetic isotope effect experiments[19, 29, 30, 32] provide a great deal of insight into protein / enzyme chemistry. However, no current experimental method is capable of providing complete structural and functional information about biomolecules. With recent advances in computational sciences, the limitations of experimental methods can be overcome to a great extent using molecular simulation methods[33]. Computational simulation methods are invaluable tools in chemistry as they can reveal information at the atomistic level, enabling explanation of the experimental observations. Because of the great complexity of biomolecules, the majority of their structural and functional information remains inaccessible to experimental techniques. Therefore computational techniques are extremely useful in understanding the structure and the functionality of biomolecules[33].

The functions of proteins are highly dependant on their flexibility and structure, which can be quite complicated[34]. Therefore understanding the dynamical behaviour (resulting from high flexibility) of a protein is of prime importance. The internal motions of proteins cover time scales of 10^{-15} to >1 s whereas length scales are from 0.01 – 10 Å[34]. The motions of proteins also cover a range of amplitudes and energies and many of these motions are critical in biochemical functions[34, 35]. The motions that are associated with catalytic activity / mechanism are localised and rapid, whereas the motions that are slow and occur on the scale of the whole protein are associated with allotropic coupling and folding.

Association of subunits takes place even at longer distance and larger time scale. Some of the important protein motions and their time scales are shown in Table 2.1 (Adopted from [34]).

Type of Motion	Spatial Extent (nm)	Time (s)
Bond-length vibration	0.2-0.5	10^{-14}-10^{-13}
Elastic vibration of globular domain	1.0-2.0	10-^{12}-10^{-11}
Rotation of solvent-exposed side chains	0.5-1.0	10^{-11}-10^{-10}
Torsional liberation of buried groups	0.5-1.0	10^{-11}-10^{-9}
Hinge bending (relative motion of globular domain)	1.0-2.0	10^{-11}-10^{-7}
Rotation of buried side chains	0.5 0.5	10^{-4}-1
Allosteric transitions	0.5-4.0	10^{-5}-1
Local denaturation	0.5-1.0	10^{-5}-10^{1}
Loop motions	1.0-5.0	10^{-9}-10^{-5}
Rigid-body (helix) motions		10^{-9}-10^{-6}
Helix-coil transitions		10^{-7}-10^{4}
Protein association	>>1.0	

Table 2.1 Protein motions and their time scales

The current experimental techniques have very limited scope in completely revealing information about the dynamic behaviour of proteins [36]. In experimental techniques, resolution of the measurement is limited with respect to space, energy and time. From experimental methods, high-resolution measurements of molecular structure are only possible for relatively rigid molecules. Also, it is impossible to analyse the different atomic interactions by experimental methods. Moreover, spectroscopic methods allow measuring relaxation times only under special circumstances and they almost totally fail to probe important processes like protein folding.

The limitations of the experimental techniques can be overcome to a very great extent by computer simulations. When sufficient theoretical models are used, they can be used to predict the structural and dynamical properties biomolecules with a degree of resolution of space, energy and time which is beyond experimental reach. Depending on the complexity of the system, either quantum mechanical or molecular mechanical methods can be used in simulations.

2.2 Quantum Mechanics in Molecular Simulations

Any molecular system can be described by a nonrelativistic, time-independent form of the Schrödinger equation as follows:

$$\hat{H}\psi_{(R,r)} = E\psi_{(R,r)} \qquad (2.1)$$

where $\hat{H}$ is the Hamiltonian for the system, ψ is the wave function, and E is the energy. In general, ψ is a function of the coordinates of the nuclei (R) and of the electrons (r). Equation 2.1 is too complex for any practical use, therefore some approximations are made[37-39].

The most important approximation is the Born-Oppenheimer approximation, in which it is assumed that the motion of the electrons is independent of that of the nuclei. This gives separate equations, one describing the electronic motion and the other describing nuclear motion[37-39].

$$\hat{H}\psi_{(r;R)} = E\psi_{(r;R)} \qquad (2.2)$$

Equation 2.2 describes the motion of the electrons and depends only parametrically on the positions of the nuclei. Therefore the nuclear positions are used as parameters that describe the molecular geometry. As a consequence of this treatment, a molecule has geometry.

The second equation then describes the motion of the nuclei as follows;

$$\hat{H}\phi_{(R)} = E\phi_{(R)} \qquad (2.3)$$

Based on the Born-Oppenheimer approximation, to calculate the energy of a molecule, it is only necessary to solve the electronic wave function and then add the electronic energy to the internuclear repulsion to get the overall energy.

The solution to the Schrödinger equation yields discrete (quantised) values (*eigenvalues*) of energy E_n and for each E_n its corresponding wave function. These wave functions are normally complex-valued. Therefore $\psi^*\psi$ is defined as the probability density, where ψ^* is the complex conjugate of ψ, and the wave functions are normalised by the requirement that the probability of finding the particle must be equal to one. This is given by Equation 2.4.

$$\int \psi^* \psi \, \vec{dr} = 1 \qquad (2.4)$$

Quantum mechanics is thus probabilistic and not deterministic; therefore, it abandons the notion of precisely defined trajectories of particles through time and space. Instead it treats them in terms of probabilities for alternative system configuration [40]. Although the original Schrödinger equation is simplified with the Born-Oppenheimer approximation to a great extent, it is still practically not possible to solve it analytically. For this reason various other assumptions are made and different models are used in today's quantum mechanical (QM) calculations.

The three main categories of QM calculations are *ab-initio*, semi-empirical (SE) and density functional (DFT). The first two employ solving of the electronic Schrödinger equation to get the electronic behaviour whereas the last method directly obtains the electronic densities[40]. In theory, these methods are capable of calculating any structural or energetical property of any molecule. These methods are used widely in force field parameterisation when empirical data is not available and serve as an invaluable tool in developing new force field parameters.

2.3 Molecular Mechanics

Application of computational methods to a system is determined primarily by the number of particles (atoms) present in the system. Therefore, simulation of large biomolecules like proteins using pure quantum mechanical methods is practically impossible. Empirical methods serve as an alternative approach to quantum mechanical methods, allowing simulation of large systems using a potential function[38].

2.3.1 Potential Function and Energy Landscape

Selection of a correct energy function is crucial in any dynamic simulation method, as inter- and intra-molecular interactions must be simulated accurately to produce chemically sensible results. In conventional molecular dynamic simulations, the Born-Oppenheimer approximation, i.e. a single nuclear coordinate, is used to represent the atom and the non-bonded interactions are expressed as a simple pair-wise additive function of those single nuclear coordinates[38]. The bonded groups of atoms are treated with two-body, three-body and four-body terms with regard to bond lengths, angles and dihedral angles respectively. In a classical force field atoms are treated as "balls" and the bonds as "springs" connecting the balls; these balls and springs are then treated with classical mechanics. A potential consists of a large number of parameterised terms and set functions to handle different components of the overall energy function.

A collection of such functions and their associated parameters is called a Force Field (FF). There are numerous force fields that are currently available for protein simulations. AMBER, CHARMM, GROMOS, and OPLS are a few of them. In this thesis the CHARMM[41] force field is used. It has the general form of a potential energy function which consists of terms for bonded and non-bonded interactions. The bonded component of the potential function holds terms for bonds, angles, dihedrals and impropers whereas the non-bonded component is made up of electrostatic and Van der Waals terms. The complete CHARMM energy function is given by Equation 2.5 and its components are shown in Figures 2.1 and 2.2.

$$\begin{aligned} V(\vec{R}) = & \sum_{bonds} k_d (d - d_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} k_\chi [1 + \cos(n\chi - \delta)] \\ & + \sum_{impropers} k_\varphi (\varphi - \varphi_0)^2 + \sum_{Urey-Bradley} k_{UB} (S - S_0)^2 \\ & + \sum_{non-bonded} \{\varepsilon_{ij} [(R_{ij}^{\min} / r_{ij})^{12} - (R_{ij}^{\min} / r_{ij})^6] + q_i q_j / e_i r_{ij}\} \end{aligned} \quad (2.5)$$

Bonding Terms

The equilibrium values of bond length, angle, dihedral angle, improper and Urey-Bradely (1-3) bond length are given by d_0 , θ_0 , χ_0 , φ_0 and S_0 respectively. The terms without the subscript zero represent their values with respect to some other configuration. Likewise, k_d , k_θ , k_χ and k_φ are the corresponding force constants in the above terms. The bond and the angle terms are modelled using Hooke's law whereas the dihedral terms are described with a cosine function. A harmonic potential is used for improper terms. The Urey-Bradley terms are not used unless fitting of specific computational results to observable vibrational spectra is required. Bonded terms are applied to all atoms which are bonded through covalent bonds in a given system. These terms represent all bonded interactions in the system and contribute to the corresponding part of the overall potential function.

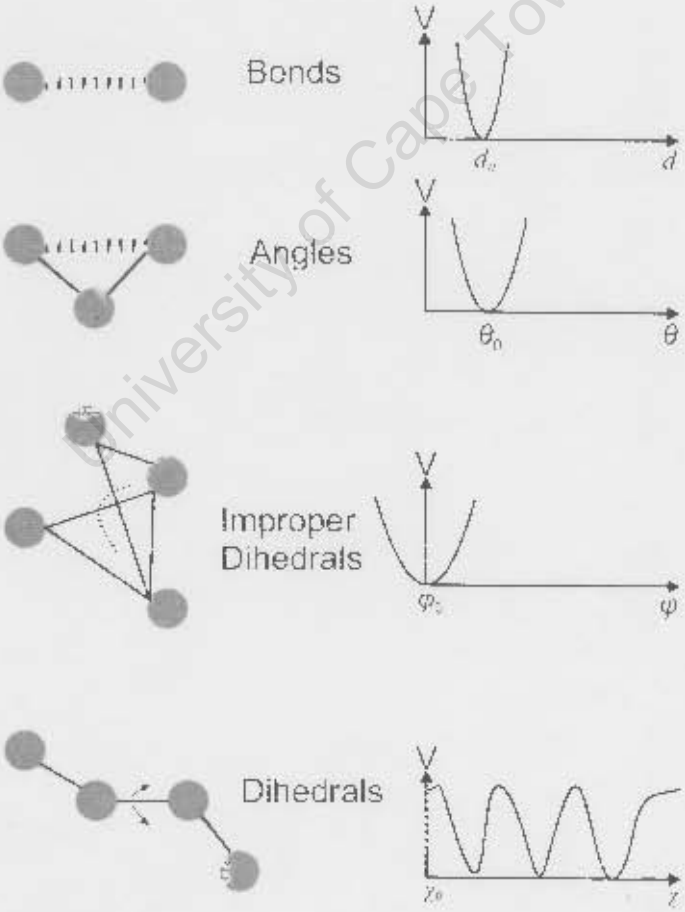


Figure 2.1 Bonding Terms of a Potential Energy Function

V= *Potential energy*

Non-bonding Terms

The non-bonded terms incorporate electrostatic and van der Waals interactions. The latter is modelled as a (12-6) Lennard-Jones interaction where in Equation 2.5, ε_{ij} relates to the Lennard-Jones well depth, R_{ij}^{min} is the distance at which the Lennard-Jones potential is minimum, q_i is the partial atomic charge of atom i, e_i is the effective dielectric constant, and r_{ij} is the distance between atoms i and j. The electrostatic terms are described as pair-wise Coulombic interactions.

Evaluation of a potential function such as the one given in equation 2.5 allows one to calculate the potential energy of a given system from a single set of atomic coordinates. When that energy is taken together with the corresponding set of atomic coordinates, it is termed a snapshot of the system.

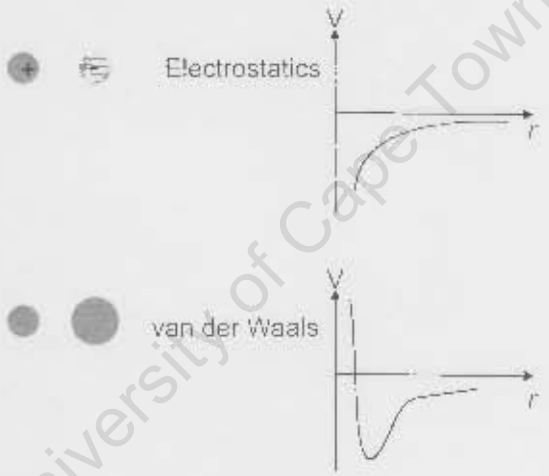


Figure 2.2 Non-Bonded Terms of a potential energy function

V = *Potential energy* r = *Inter-particle distance*

2.3.2 Energy Minimisation

The energy landscape of a biomolecule has an enormous number of minima, i.e. stable conformational sub states. The principle goal of energy minimisation (EM) is simply to find the local energy minimum, i.e., the bottom of the energy well occupied by the initial conformation. This concept is schematically represented in Figure 2.3.

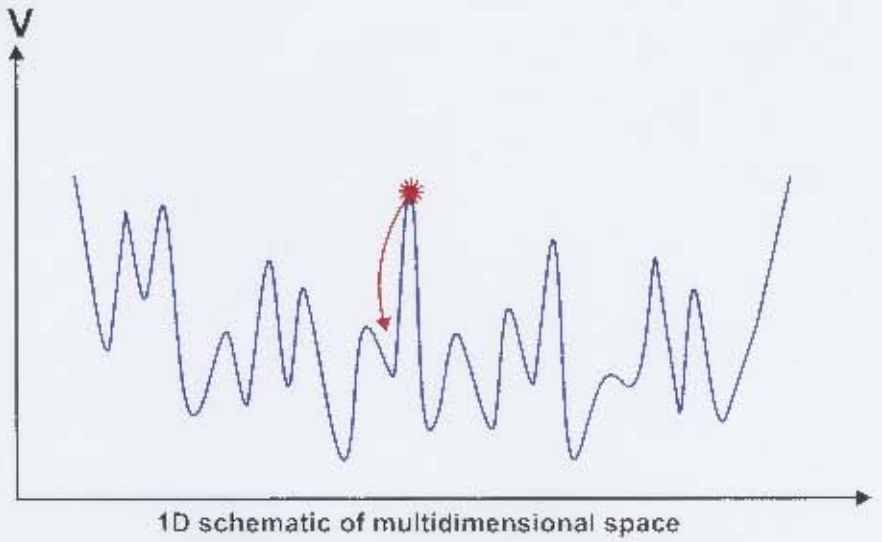


Figure 2.3 Schematic representation of energy minimisation

On the energy surface there may be a very large number of minima. The minimum with the lowest energy is termed the global energy minimum. At any other point away from these minima, the corresponding structures have higher energies. Therefore, in order to represent a given system successfully, its geometries corresponding to the minima in the energy surface must be identified. This procedure is referred to as energy minimisation. The minimisation problem is formally stated as follows: given a particular potential energy function V, which depends on N independent variables of r, $r=r1, r2, r3, ..., r_N$; the task is to find the values for each of those variables, termed r_{min}, for which V has its global minimum[38]. Evaluation of the first and the second derivatives of the function V with respect to each of the variables is a way to identify the minima, where the value of the first derivative is equal to zero and the second derivatives are all positive(Equation 2.6):

$$\frac{\partial V}{\partial r_N} = 0 ; \quad \frac{\partial^2 V}{\partial\, r^2{}_N} > 0 \qquad (2.6)$$

Energy minimisation methods can be very useful for correcting an unrefined molecular structure with bond angles and lengths distorted from their respective minima or with steric clashes between atoms, therefore EM is routinely applied to molecular systems prior to any simulations. However, it is impossible to locate the global minima of the energy landscape of a biomolecule when there are a few hundred atoms in the system.

Although energy minimisation methods may be used to refine molecular structures efficiently; they are totally inadequate for sampling conformational space and they cannot provide any time-dependent properties of the system.

Energy minimisation is done by using an appropriate minimisation algorithm. The most popular energy minimisation methods include those that use derivatives of various orders. The steepest descent and the conjugate gradient algorithms use the first order derivative whereas the Newton-Raphson algorithm uses the second order derivatives. Also, there exist non-derivative minimisation algorithms such as the Simplex method and the Sequential Univariate method.

2.4 Methods for Simulating Biomolecules

Simulations of biological macromolecules are performed using various methods that can handle the size and complexity and can reveal valuable information about them. As explained in the previous sections, pure quantum mechanical approaches cannot be used due to the size constraints. The methods that are currently available for biomolecular simulations are either based purely on molecular mechanics or are a combination of quantum mechanical and molecular mechanical approaches. The subsequent sections of this chapter will describe those simulation methods with an extended focus on Molecular Dynamics methods as these are extensively used in this thesis.

As mentioned previously, computer simulations of biomolecules try to access information about the system of interest in terms of space, time and energy. Therefore, simulation methods of biomolecules can be categorised into three main classes. These are: simulations that are performed to obtained structural information, simulations used to obtained energetically important results and lastly simulations that are performed to obtained time-evolutionary information[33]. Therefore, a selection has to be made between these methods based on the desired information that is expected to be obtained from the simulation.

When the correct simulation method has been chosen, it is necessary to select an appropriate model to represent the molecular system of interest. The most commonly used models are; the all-atom model, the united- atom model, the mesoscopic and the coarse-grain model. In the all-atom model, as the name implies, atoms are treated individually in the simulation. The united-atom model uses a "united-atom", which is a particle that incorporates a group of atoms but can approximately represent the molecular mechanical properties of the group on a scale of size that is larger than atomic scale. The last two models (used in protein modelling) are "reduced" models in which each amino acid is represented by only a few interaction points. Therefore, depending on the desired information to be obtained from a simulation, selection of an appropriate model is vital.

2.5 Molecular Dynamics (MD)

Molecular Dynamics (MD) simulation methods allow one to calculate the time-dependent behaviour of a molecular system. As such, it has become an invaluable tool in determining time-evolutionary structural and thermodynamic properties of molecular systems and these methods are routinely used to investigate properties of bio-macro molecules including proteins.

In a molecular dynamics simulation, the macroscopic properties of a system are explored through microscopic simulations. The connection between microscopic simulations and macroscopic properties is made via statistical mechanics. Molecular dynamics simulations provide the means to solve the equation of motion of the particles in the system, which enables one to relate macroscopic properties to the distribution and motion of the atoms and molecules of the N-body system as governed by the rules of statistical mechanics. Therefore MD is a method based on statistical mechanics and classical physics [42-44]. In MD simulations a reasonable guess is made of the interactions between molecules, and it is attempted to obtain `exact' predictions of bulk properties. This is done by means of a potential energy function (described in section 2.3.1) that models the basic interactions. In a Force Field (FF), i.e. a potential energy function, the energy, E, is a function of the atomic positions, R, of all the atoms in the system.

There are two main types of closely related molecular dynamic simulation methods; Molecular Dynamic (MD) Simulations and Stochastic Dynamic (SD) simulations. Although the term "Molecular Dynamics" is commonly used for both MD and SD, the principal equations that are solved in these two methods are completely different. These two methods are discussed in the following sections.

In Classical Molecular Dynamics, atoms and bonds are treated as 'balls' and 'springs' and Newton's equations of motion are used to calculate *R(atomic positions);* hence the energy of the system. According to Newton's law of motion, the force F exerted on particle i is given by equation 2.7,

$$F_i = m_i.a_i = m_i.\frac{dv_i}{dt} = \frac{m.d^2r_i}{dt^2} \qquad (2.7)$$

where m_i is the mass of particle i and a_i is the acceleration of particle i. The force is related to the potential energy E according to equation 2.8.

$$F = -\nabla_i E \qquad (2.8)$$

A combination of Equations 2.7 and Equation 2.8 yields:

$$-\frac{dE}{dr_i} = m_i \frac{d^2r_i}{dt^2} \qquad (2.9)$$

Equation 2.9 relates the derivative of the potential energy to the changes in position as a function of time. One can calculate a trajectory for a given system by solving equation 2.9, where trajectory is defined as a collection of configurations of the system as it evolves with time. Calculation of a trajectory requires the initial positions and the initial distribution of velocities for all atoms in the system. The following set of equations show how a trajectory is calculated in a MD simulation. In Equations 2.10 – 2.14 a is the acceleration, v is the velocity, t is the time and x is the position and the subscript 0 indicates the starting values.

If the acceleration is constant,

$$a = \frac{dv}{dt} \qquad (2.10)$$

After integration of Equation 2.10

$$v = at + v_0 \qquad (2.11)$$

and

$$v = \frac{dx}{dt} \qquad (2.12)$$

After integration of Equation 2.12

$$x = v.t + x_0 \qquad (2.13)$$

From Equation 2.11 and 2.13,

$$x = a.t^2 + v_0.t + x_0 \qquad (2.14)$$

Equation 2.14 gives the value of x at time t as a function of the acceleration, a, the initial position, x_0 , and the initial velocity, v_0. Equation 2.9 gives the potential energy of the system as a function of time and position. If the solutions to equations 2.9 and 2.14 can be obtained for all the atoms in the system at all times, then all the positions (hence the overall configuration of the system) and the potential energy at all times can be obtained. When those two entities are known any thermodynamic property of the system can be calculated.

The equations of motion are deterministic, i.e., the positions and the velocities at time zero determine the positions and velocities at all other times, t. The initial positions can be obtained from experimental structures, such as the x-ray crystal structure of the protein or the solution structure determined by NMR spectroscopy.

The initial distribution of velocities is usually determined from a random distribution with the magnitudes conforming to the required temperature and corrected so that there is no overall momentum, i.e.,

$$P = \sum_{i=1}^{N} m_i V_i = 0 \qquad (2.15)$$

Once the initial conditions are decided, Newton's equations must be integrated in order to generate the results. This can be done using one of many available "integrators" i.e. integrating algorithms. Integrating algorithms are briefly discussed in section 2.8.

2.6 Langevin Dynamics

Stochastic terms are included in Langevin dynamics (LD) to approximate the effects of degrees of freedom that are neglected in a regular MD simulation. It employs the Langevin equation which has two additional terms, as an alternative to Newton's second law. The first term is a frictional function that intends to represent the frictional drag experienced by solute molecules in a solvent. The second term represents a random force that is applied to mimic the random impulses that would be expected from both the solvent and any coincident solute molecules. For a given atom i, the Langevin equation is given as,

$$m_i \frac{d^2 r}{dt^2} = F_i(r) - \xi_i \frac{dr}{dt} + R_i(t) \qquad (2.16)$$

where $F_i(r)$ is the force F exerted on particle i (Section 2.5), ξ is the friction coefficient, and $R_i(t)$ represents the random forces experienced by the atom. The relationship between ξ and $R_i(t)$ regulates the temperature of the simulated system. When ξ=0, Langevin dynamics is equivalent to conventional MD. When $\xi > 0$, the random impulses felt by the system can support to propagate barrier-crossing motions. Therefore, improved conformational sampling can be obtained from Langevin dynamics.

2.7 QM and QM/MM Methods

As discussed previously, molecular mechanics-based methods such as MD and LD serve as successful alternatives to pure quantum mechanical methods in simulating biomolecules. Although, these methods are capable of producing accurate data (structural, functional, energetical etc.) of the system of interest, they still suffer some major drawbacks. The inability to sample all available phase space (generally referred to as sampling problem), and to study reaction mechanisms (bond forming and bond breaking) are among the major problems in conventional MD/LD simulations.

Methods that use a combination of both quantum mechanics and molecular mechanics are becoming more popular due their high accuracy and increasing feasibility. These methods are generally referred to as Quantum Mechanical – Molecular Mechanical (QM/MM) methods or hybrid methods. The hybrid methods allow one to take advantage of the quantum mechanical approach in terms of accuracy and that of molecular mechanical approach in terms of efficiency.

In QM/MM methods, a small part (the most important region / regions) of a large system such as a protein is treated with quantum mechanics while the rest of it is treated with molecular mechanics. Partitioning of a protein to different regions in a QM/MM simulation is given in Figure 2.4. This allows one to obtain information at an electronic level for the atoms that are treated quantum mechanically, i.e. it allows simulation of the reaction mechanisms which is impossible with classical mechanical methods. However, the main disadvantage of these methods is still their computational cost. Because of this, most of these types of calculations are done using semi-empirical methods rather than *ab-initio* /MM methods [45].

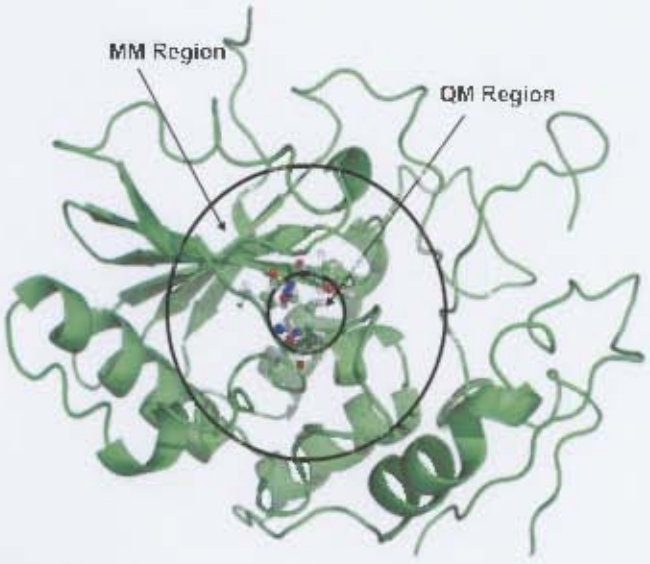


Figure 2.4 Partitioning of a protein in a QM/MM simulation

In QM/MM methods, the Hamiltonian ($\hat{H}$) of a system is composed of three Hamiltonians and is given by: Equation 2.17.

$$\hat{H} = \hat{H}_{QM} + \hat{H}_{MM} + \hat{H}_{QM:MM} \tag{2.17}$$

where $\hat{H}_{QM}$ corresponds to the quantum part, $\hat{H}_{MM}$ to the classical part and $\hat{H}_{QM:MM}$ to the interaction between them. The energy of the system and the state function are then obtained by solving the wave function of the system with the correct Hamiltonian (based on Equation 2.17). This requires a correct definition of the term $\hat{H}_{QM:MM}$ for the interactions between these two regions. There are different QM/MM models based on how they treat the $\hat{H}_{QM:MM}$ term. The quantum mechanical methods and the molecular mechanics methods were discussed in sections 2.2 and 2.3 respectively.

2.8 Integration Algorithms

As shown in the previous sections, calculations of the MD/ LD components require one to integrate the equation of motion for all interacting particles in a given system. Since the potential energy is a function of the atomic positions of all the atoms in the system, this complexity does not allow solving the equations of motion analytically.

Hence, they must be solved numerically. An appropriate integration algorithm among many available can be employed to accomplish this requirement.

In integration algorithms, finite difference methods are used. This is done by discretising time on to a finite grid. A time step (Δt) is then defined as the distance between two consecutive points on the grid. If the initial positions and velocities are known, an integration algorithm can give the same quantities at a later time. Iteration through this procedure enables one to follow the time evolution of the system for long periods of time. Among many available, the following are the most popular integration algorithms used in MD: Verlet algorithm[46], leap-frog algorithm[47] velocity Verlet [48] and Beeman's algorithm[49].

2.9 Simulation Environment

A wide range of experimental conditions can be simulated by using MD simulations. Therefore, a selection has to be made about the conditions prior to a calculation. In making this decision, the following entities are of prime importance. They are; Simulation ensemble, solvation (implicit or explicit), boundary conditions and long range interactions. These are discussed in sections 2.9.1- 2.9.3.

2.9.1 Solvation

It is not possible to simulate the full physiological conditions of a protein in a computer simulation. However efforts are made to incorporate as many as possible of those conditions in a simulation and solvation is one of the key phenomena. As most proteins exist, at least partially, within an aqueous environment it is crucial to simulate the solvation correctly. There are two main models to incorporate the solvents in a simulation; these are implicit models and explicit models.

2.9.1.1 Implicit Solvation

A large amount of computer simulation time is taken to evaluate the solvent-solvent effects. However, the solvent effects can not be totally ignored. The implicit solvent models are originally designed to address the above issues and there are numerous implicit water models that have been developed[50-52].

These models are called continuum models as they represent the solvent as a continuum medium instead of individual solvent molecules.

Although these models have proved their ability to produce better results in certain types of calculations [53] (e.g. solvation free energy calculations), due to the fact that they facilitate undesirable conformational changes in proteins, these methods are not generally used in protein simulations.

2.9.1.2 Explicit Solvation

The solvent is represented as individual or explicit molecules in this model in contrast to implicit models. Simulation of a system with only a thin layer of water / solvent around it can overcome the majority of the problems of pure implicit models. Currently, there is a wide range of explicit water models available and commonly parameters of these water models are adjusted to reproduce the enthalpy of vaporisation and density of water. Some of the more common and popular explicit water models are TIP3P[54], TIP4P[54], TIP5P[55], and SPC/E[56]. These models can be classified based on the number of points used to define the model (atoms plus dummy sites); rigidity / flexibility, and polarisation effects.

The TIP3P water model is used in this thesis owing to its proven ability to mimic solvent effects in protein simulation and, more imprtantly, CHARMM's inherited usage of this model. All CHARMM force fields (eg. proteins, nucleic acids, lipids) have been parameterised with respect to TIP3P. It is the simplest model of explicit water and has three sites of interactions corresponding to the three atoms of the water molecule. The partial positive charges on the hydrogens are balanced by an appropriate negative charge on the oxygen atom and the van der Waals interactions between two water molecules are calculated using a Lennard-Jones function with a single point of interaction per molecule which is centered on the oxygen atom. There is no interaction calculated between the hydrogen atoms. TIP3P has a rigid geometry and the original model has been modified by placing Lennard-Jones parameters on hydrogen in CHARMM.

2.9.3.1 Periodic Boundary (PB) Conditions

In the PB method, a periodic cell is replicated throughout the space to form an infinite lattice. Even though a cubic cell (often called a "box") is often used, there are many other geometrical shapes of periodic cells. Throughout the simulation, if a molecule in the central cell moves, its image in the neighbouring cells moves exactly the same way and in the case of a molecule leaving the original cell during the simulation, one of its images will enter the central cell through the opposite face. Therefore, there are no walls or surface molecules in the central cell. However, the concept of periodic imaging used in this method consists of an infinite number of terms for a given particle that must be solved in order to evaluate the potential energy function. This practically impossible target is achieved using the techniques called Minimum Image and Truncation of the Potential (see section 2.10).

2.9.3.2 Non-periodic Boundary Conditions

With increased computer power, it is possible to explicitly include water as a "skin" surrounding the system of interest. This idea is implemented in methods based on liquid droplets / stochastic boundaries and van der Waals cluster methods. Owing to their models, these methods inherently contain boundaries. For this thesis stochastic boundary conditions (SB) are used. This method is particularly useful in investigating only a particular region such as the binding site in a ligand-binding study. This allows much of the system that would otherwise be simulated to be excluded.

SB uses elements of both Langevin dynamics and MD. In SB the system of interest is divided into three main regions called reaction region, buffer region and reservoir region. The reaction region is the site of interest and the atoms in this region are subjected to the full simulation method. The buffer region is simulated with Langevin dynamics, used to eliminate undesired degrees of freedom by imposing boundary and stochastic forces on the system. This region manages any local fluctuations in energy, conformation or density that take place in the reaction region. The atoms in the reservoir region are kept fixed.

2.10 Truncation of the Potential

As the number of atoms in a system increases, the number of non-bonded terms that need to be evaluated increases as the square of the number of atoms. Hence, evaluating all these interactions become computationally very expensive and practically nearly impossible. This problem can be dealt with by using a non-bonded cutoff. When a non-bonded cutoff is employed, the interaction energy between the atom pairs further apart than the cutoff distance is set to zero.

However, cutoffs introduce immediate discontinuity of both potential energy and the forces near the cutoff value. In order to overcome this problem, a "smoothing" technique is applied. The most widely used methods in this category are the switching and shifting functions.

2.11 A Protocol for Performing a Molecular Dynamics Simulation

A typical molecular dynamics simulation has the following fundamental steps.

1. Preliminary preparation

A molecular structure with the coordinates for all atoms in the system is required for a molecular dynamics simulation. The molecular structure can be built using a molecular modelling software package or can be directly obtained from x-ray crystallography experiments. Especially in the case of proteins, structures that are obtained from crystallography require further refinement. The initial preparation of a protein is discussed in more detail in section 2.12.

2. Minimisation

Usually, the starting structures (either from modelling or from experiments) contain steric overlaps i.e. bad contacts. Therefore, they are normally far from a realistic conformation of the actual molecule. In order to overcome this problem an energy minimisation must be performed to obtain a reasonable starting geometry. This is explained in detail in section 2.3.2.

3. Heating

A minimised structure represents the molecule near to absolute zero and therefore is not suitable to simulate a system at normal working conditions. The temperature must be increased, and this is done in a simulation by assigning random velocities to molecules to reproduce a Gaussian distribution.

4. Equilibration

The simulating system must stay steady in terms of the temperature and the structure in order to represent a realistic system. In a MD simulation it is done by allowing the system to evolve spontaneously. A desired temperature (including all other simulation parameters) and an appropriate integrating algorithm must be used in phase. The length of this phase is system-dependent. A small and simple system might take ten to a few hundred picoseconds to equilibrate whereas a larger complicated system (like a protein) will take a few hundred picoseconds to a few nanoseconds. Generally, when the statistical properties of the system become independent of time it is considered to be equilibrated. Time series of the total energy of the system or the root mean square deviation (RMSD) of the structure are very helpful in checking whether the system has reached equilibration or not. Generally a protein system with a few thousand atoms might need an equilibration period of not less than 1-5 ns.

5. Production

The actual dynamics are performed in the production phase, using the equilibrated structure as the starting point. The trajectories are generated in this phase following the time evolution of the system. The length of the production also depends on the size of the system and required data.

2.12 Initial Preparation of Proteins

2.12.1 Structure Refinement

In most cases of protein simulations, the starting coordinates are obtained from x-ray crystallographic experiments. However, the x-ray structures are normally distorted to a certain degree due to the "crystal environment" and related symmetry issues.

Some of these issues are atomic clashes (normally referred to as "bumps") and chi-flipping i.e side chains having a wrong chi angle and missing chains. These issues must definitely be solved before any simulation is performed on the protein. There are numerous applications available for this purpose and for this thesis the WHATIF[57] program was used.

2.12.2 Determination of Protonation States

Uptake and release of protons by amino acids is one of the more frequent chemical reactions that occur in protein-water solutions. Specific moieties of some of the amino acids in a protein which can show the above phenomenon are referred to as protein titratable groups (TG). There are two categories of titratable groups: acids and bases. The C-terminal of a protein and the amino acids ASP, GLU, CYS, SER, TYR, and THR are categorised as acids. Bases are the N-terminal, HIS, LYS and ARG. Acidic TG s have an overall neutral charge when they are protonated and the basic TG s are positive when they are protonated.

Therefore, having the correct protonation states for those amino acids is crucial in protein chemistry and simulations. However, with the current crystallographic techniques, it is not possible to obtain the coordinates of the hydrogen atoms (protons) due to the low electron density around them. This then becomes an issue for computer simulations when structures are obtained from x-ray crystallographic results. In order to determine the protonation states of the TG s, numerous methods have been developed. In almost every method, this is done by calculating the pKa values of the TGs. One of the most popular ways of calculating the pKa is by employing electrostatic calculations using the finite difference Poisson-Boltzmann (FDPB) method. Calculating pKa values using free energy methods is also not uncommon.

For this thesis, the WHATIF program was used to determine the protonation states of the protein. WHATIF uses free energy methods to perform pKa calculations. As with some of the other computational chemistry packages, this program uses the concept of optimizing the hydrogen bonding network as the basis for protonation state determination.

WHATIF treats the pKa value of a titratable group as a measure of the free energy difference between the neutral and charged state of the group and tries to calculate the pKa value by calculating the free energy difference between the charged and neutral state of that group in the protein. It performs these calculations in three steps. These are the calculation of the desolvation energy, the background interaction energy and the pair-wise interaction energy between the titratable groups. The pKa curves of the amino acids generated by these calculations can then be used determine their protonation state.

However, none of the existing pKa determination methods is perfect. This inadequacy in pKa calculations is due to an incorrect description of the protein in the calculations, because, pKa calculations use the crystal structures as the source of coordinates for the protein. The crystal symmetry induces structural changes in the protein, and thereby causes some pKa values to be shifted compared to their value in solution. Therefore it is not surprising that the pKa values calculated from a crystal structure will differ from the pKa values measured in solution by NMR, and determination of accurate protonation states still remains challenging.

2.13 Parameterisation

Specification of the terms in the potential energy function for a chosen molecular system is termed as parameterisaion. A widely accepted protocol for parameterisation of novel molecules for a force field is given below. In this protocol a self-consistent approach is employed and all parameters are introduced based on a "Interaction Triad" (Figure 2.5) in which all possible types of solute-solvent interactions are taken into account.

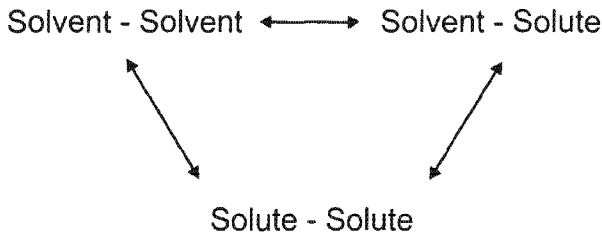


Figure 2.5 The Interaction Triad

When the atom types are specified, depending on the chemical environment, it is then necessary to define all the other bonding (force constants, equilibrium distances, angles, dihedrals etc) and non-bonding (partial atomic charges and Lennard-Jones parameters) parameters. Parameters are obtained from various types of resources. The force constants for stretching and bending (hard degrees of freedom) are usually obtained from experimental vibrational data. Parameters such as torsion, van der Waals and electrostatic terms are difficult to obtain experimentally. Therefore, theoretical quantum mechanical calculations are performed on model compounds to generate those parameters. These parameters, in conjunction with the partial atomic charges, are then fitted to reproduce the calculated torsional barriers with the relative energies of the different conformations.

The extent of parameterisation depends on the quality of data that is going to be generated using the new parameters. Usually, the extent of parameterisation is classified as minimal parameterisation and maximal parameterisation. Minimal parameterisation is done by analogy, where the new parameters are adopted from existing parameters. The parameters obtained from this are considered as starting parameters and usually require further refinements for high accuracy data. However, this method is used when the target molecule is very similar to any of the topologies in the existing force field and when there is a time constraint for intensive parameterisation. It is usually necessary to determine the partial atomic charges for the target molecule even with the minimal parameterisation, as they are highly dependent on the chemical environment.

Partial atomic charges can be obtained from quantum mechanical calculations. Either Mulliken charges or any other charge-fitting scheme (e.g. Merz-Kollman) is used to obtain the initial charges and they are then fitted to the force field by scaling with an appropriate scale factor.

The maximal parameterisation is done by refining new parameters in a recursive fashion to fit the target data. This concept is illustrated in Figure 2.6. Although the accuracy of the generated data is very high, this is highly time-expensive and requires a significant amount of target data.

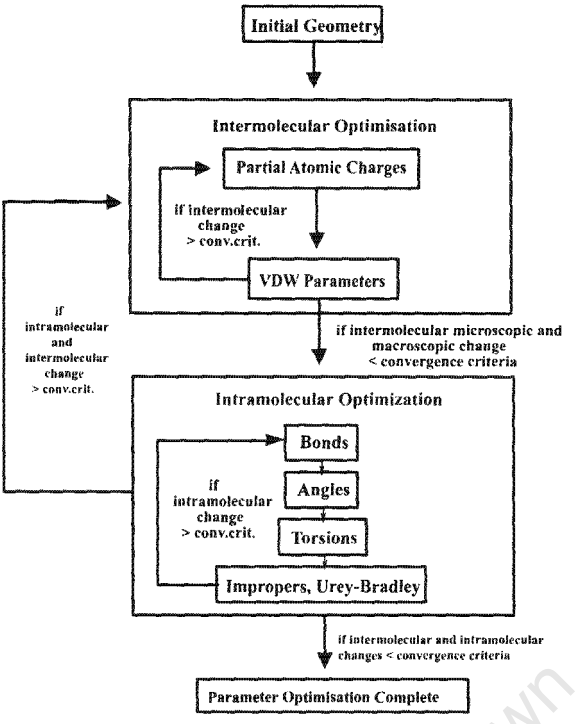


Figure 2.6 Parameterisation Protocol

For this thesis, parameterisation was done by analogy and the partial atomic charges were determined using the Merz-Kollman (MK) method. The determination of partial atomic charges using MK method is described below.

2.13.1 Merz-Singh-Kollman (MK) Scheme

The scheme was proposed by U. C. Singh and P. A. Kollman[58]. The method involves fitting the atomic charges to reproduce the molecular electrostatic potential (MEP) at a number of points around the molecule. Initially the MEP is calculated at a number of grid points located on several layers around the molecule, and the layers are constructed as an overlay of van der Waals spheres around each atom. The points that are located inside the van der Waals volume are then discarded. Four layers are constructed and the radius of the smallest layer is obtained by scaling all radii with a factor of 1.4. and the next three layers with the scale factor of 1.6, 1.8, and 2.0. After evaluating the MEP at all valid grid points located on all four layers, atomic charges are derived that reproduce the MEP as closely as possible. Generation of layers is schematically given in Figure 2.7.

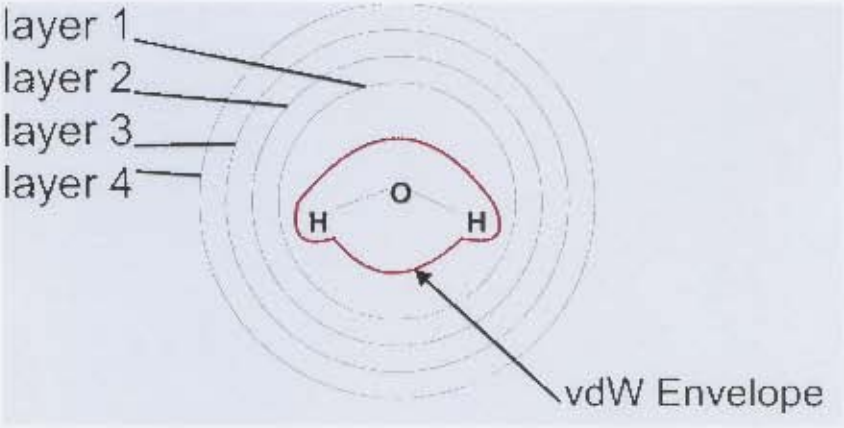


Figure 2.7 Schematic representation of MK charge fitting methods

A choice between the methods is made based on the required data, size of the system, etc. For this thesis , calculation of partial atomic charges was performed with DFT calculations with the B3LYP[59] functional with the 6-31G* basis set. The B3LYP function incorporates an electronic exchange-correlation and is therefore considered to be a "fairly good" model.

2.14 Simulation Analysis:

2.14.1 Time Series

A time series is defined as a collection of observations of well-defined data items obtained through repeated measurements over time at uniform time spaces. In general a time series can be divided into three components. They are the trend (long term direction), the seasonal (systematic, short-term movements) and the irregular (unsystematic, short -term fluctuations).

In a computer simulation (dynamics simulation) of a molecular system, a set of trajectories is generated. These trajectories can be used to create time series of the desired entity. Time series of energy of a system, geometrical parameters such as bond length, angle, dihedral, and distances, root mean square deviation (RMSD) of structure and interaction energies are among the most widely used analysis tools in simulations.

Auto-correlation functions and cross-correlation functions are special types of time series that are used to study correlations between two sets of data. Time series are very helpful in understanding the underlying context of the data points (their origin, what generated them?), or in making predictions and an appropriate analysis method or forecasting method is used to do this.

The most obvious information that can be obtained from a time series is the behaviour of the monitored entity over time and this can serve as a very valuable tool in molecular simulations to make early decisions such as whether the system is equilibrated or not, is temperature steady etc.

2.14.2 Hydrogen Bonding

Hydrogen bonding is an attractive intermolecular force that exists between two partial electric charges of opposite polarity, where on one side hydrogen is attached to a strongly electronegative heteroatom, such as oxygen, nitrogen or fluorine, which is called the hydrogen-bond donor, and on the other side there is a heteroatom with a lone pair of electrons to accept the positive charge created on hydrogen; this is called an acceptor. Hydrogen bonds can be intramolecular or intermolecular. Hydrogen bonds are stronger than any other intermolecular interactions and therefore play a significant role in molecular association, especially in ligand-protein binding.

The hydrogen bond is not simply an attraction between point charges but it possesses a degree of orientational preference. Also, the hydrogen bond can be shown to have some of the characteristics of a covalent bond, and the degree of covalency tends to be higher when acceptors bind hydrogens from more electronegative donors. Analysis of hydrogen bonds in a molecular simulation is standard practice and it reveals useful information about molecular interactions. Usually, the criterion used for hydrogen bonding depends on the simulation methodologies used.

This is a consequence of the fact that hydrogen bonding can only be completely identified with the aid of quantum mechanics. Therefore when QM is not used, i.e. in force field based simulations, an appropriate definition of hydrogen bond is mandatory.

An example of a definition of the hydrogen bond in force field methods is given below, and this criterion was used for this thesis. A hydrogen bond exists when,

1. The donor-hydrogen-acceptor angle is between 120° and 180°
2. The distance between an acceptor and donor is less than 2.4 Å

2.15 Application of Computational Chemistry Methods in Drug Discovery

Computational chemistry methods are widely used in the modern pharmaceutical industry. Computer Aided Drug Design (CADD) is a valuable tool in rational drug design protocol. The range of application of computational methods is vast. Modern computational techniques are capable of providing information on solvation effects, free energies of binding, molecular motion, reaction dynamics and molecular properties. The methods that are used include molecular mechanics, molecular dynamics, quantum mechanics, and hybrid methods of quantum and molecular mechanics.

Two distinct areas are identified in CADD. The first category is where detailed molecular structure of the target macromolecule, the drug receptor, is known, and this is usually referred to as rational drug design or structure-based drug design. The other category is when the target receptor binding site has properties which can only be inferred from knowledge of the variable activity of otherwise similar molecules, and this is termed combinatorial drug design.

In rational drug design a new molecule (drug) that can specifically interact with the biomolecules is modeled using computational tools. The most obvious scaffolds for a novel drug molecule are the original substrate of the targeting enzyme and the products of its catalytic mechanism.

However, based on the transition state theory, it is possible to design a new drug / inhibitor based on the transition state structure, i.e. a transition state analog. The most common approach to design a TS analogue would be to compute the energy profile of a biochemical transformation which it would be desirable to inhibit and then locate the transition state or intermediate and subsequently create a stable mimic of these unstable transients. Such a mimic is recognised by the enzyme responsible for catalysing the reaction and it would hence act as an inhibitor. Identification of the TS can be done by employing QM or QM/MM methods. Identification of the TS can be quite challenging when the reaction mechanism is unknown. The modern computational chemistry techniques can be used to study the reaction dynamics (see chapter 5) and therefore, has invaluable capabilities of enhancing the process of rational drug design.

A few examples of how computational methods / tools are used in the field of medicinal chemistry is given below.

- Construction and visualisation of 2D and 3D structures
- Obtaining molecular dimensions and properties
- Conformational analysis / identifying active conformations
- 3D pharmacophore identification
- Docking
- Automated screening of data bases for lead compounds
- Receptor mapping / construction of a receptor
- QSAR studies

CHAPTER 3

Free Energy Calculations

3.1 Introduction

Free energy is one of the most important phenomena in chemistry. The tendency of a molecular system to change or to react is determined by its free energy and the probability of finding a system in a given state with respect to another is determined by its free energy and the free energy difference between the two states [60]. The energetical properties of a given system can be determined in different ways. Free energy calculations in general refer to a class of simulations that are related through classical statistical mechanics equations, enabling the free energy of a system to be determined from its microscopic descriptors[60]. Those thermodynamic properties can also be expressed in terms of statistical mechanical ensemble averages, where the ensemble averages are approximated by calculated ensemble or time averages from molecular simulations[61].

There are two widely used classes of empirical simulation methods available to generate ensemble averages. They are Molecular Dynamics (MD) and Monte Carlo (MC) methods (Chapter 2 of this thesis). Each method uses an analytic potential energy function dependent on the potential energy to generate an ensemble of configurations. In this thesis the CHARMM[41] program was used to perform the free energy calculations and the potential energy function used in the CHARMM program is given in Equation 3.1.

$$\begin{aligned} V(\vec{R}) = & \sum_{bonds} k_d (d - d_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} k_\chi [1 + \cos(n\chi - \delta)] \\ & + \sum_{impropers} k_\varphi (\varphi - \varphi_0)^2 + \sum_{Urey-Bradley} k_{UB} (S - S_0)^2 \\ & + \sum_{non-bonded} \{\varepsilon_{ij} [(R_{ij}^{\min} / r_{ij})^{12} - (R_{ij}^{\min} / r_{ij})^6] + q_i q_j / e_l r_{ij}\} \end{aligned} \tag{3.1}$$

Molecular Dynamics generates a time-dependent thermodynamic ensemble from which the desired averaged quantities are calculated. In comparison, Monte Carlo generates a thermodynamic ensemble using a series of random moves, which are evaluated based on energy criteria[62]. Therefore, the main difference between MD and MC is the lack of "time evolution" of the latter method. As a result, MC has more limitations in overall sampling when compared to MD. However, both methods are unable to traverse important regions of the phase space within a reasonable amount of simulation time.

For a system with a constant number of particles, volume and temperature its free energy is expressed as a Helmholtz function *(A)* and is given by,

$$A_{(N,V,T)} = E_{(N,V,T)} - TS_{(N,V,T)} \tag{3.2}$$

where, *E*, *T*, and *S* correspond to enthalpy, absolute temperature and entropy respectively. In statistical mechanics, the partition function (*Z*) of the canonical ensemble (constant *NVT*) is related to the free energy *(A)* of the system by,

$$A_{(NVT)} = -k_B T \ln Z_{(NVT)} \tag{3.3}$$

$$= -k_B T \ln[(h^{3N} N!)^{-1} \int\int e^{-H(p,r)/k_B T} dp\, dr \tag{3.4}$$

where *h* is the Planck's constant, k_B is the Boltzmann constant and T is the absolute temperature. *H* is the classical Hamiltonian, given by,

$$H(p,r) = \sum_{i=1}^{N} p_i^2 /(2m_i) + U_{(r)} \tag{3.5}$$

where *r* represent the coordinates ($r_1, r_2, .. r_N$) , *p* is the conjugate momenta ($p_1, p_2, .. p_N$) , *N* is the number of particles , *m* is the mass of the particle and $U_{(r)}$ is the interaction function.

According to equation 3.4, the free energy of a classical system is given by the double integral of $e^{-H(p,r)/k_BT}$ over all possible values of p and r, which defines the volume of the phase space accessible by the system. However, as mentioned before both MD and MC are incapable of sampling the total accessible phase space of the system. Therefore, it is not possible to estimate the absolute free energy of a system by finding a numerical solution to equation 3.4 classically.

3.2 Calculating the Free Energy Difference

Although calculating accurate absolute free energies is nearly impossible in practice, it is possible to calculate the free energy difference between two closely related states. Currently there are a number of methods available for this purpose. These simulation methods relate the free energy difference (ΔG) between two states of the system to a thermodynamic ensemble average that is reliant on the potential energy properties of those states[63]. Therefore, ΔG can be expressed as,

$$\Delta G = f(\text{ensemble averaged function of the potential energy}) \tag{3.6}$$

There are three commonly used methods to calculate the free energy differences. They are: Thermodynamic Integration (TI), Slow Growth (SG) and Free Energy Perturbation (FEP). In all these methods, the so-called coupling parameter approach is used. When the free energy difference between two closely related states (A and B) is calculated the Hamiltonian is given as a function of the coupling parameter λ. When $\lambda = \lambda_A$ the Hamiltonian corresponds to the state A where, $H(p,r,\lambda_A) = H_A(p,r)$, and when $\lambda = \lambda_B$ the Hamiltonian corresponds to the state B where, $H(p,r,\lambda_B) = H_B(p,r)$. Since the Hamiltonian can be written as a function of λ, so the partition function can be given as a function of λ, as given by Equation 3.7,

$$Z_{(NVT;\lambda)} = (h^{3N}N!)^{-1}\int\int e^{-H(p,r;\lambda)/k_BT}\,dp\,dr \tag{3.7}$$

As a consequence of equation 3.7, the free energy given by equation 3.3 also becomes a function of λ and it is given by,

$$G_{(NVT;\,\lambda)} = -k_B T \ln Z_{(NVT;\,\lambda)} \qquad (3.8)$$

According to equation 3.8, the change in free energy (ΔG) also becomes a function of λ. Any of the free energy calculation methods mentioned above (TI, SG or FEP) can then be used to calculate the change in free energy for a given process. These methods differ from each other due to their unique approaches of computation of relative free energy by solving the Hamiltonian as a function of λ. All these methods are based on the simple fact that the free energy is a state function and is independent of the path or the simulation method. One has to make the decision of making the choice between these methods based on the system of interest. These methods will be discussed in the following sections of this chapter and the FEP method will be discussed in more detail as it was intensively used for this thesis.

3.3 Thermodynamic Integration (TI)

In TI, free energy is calculated by integrating along the path that takes the system from one thermodynamic state to another. The Hamiltonian (H) of the system is expressed as a function of the coupling parameter (λ). The Hamiltonian of the initial state is given by $H_{(\lambda=0)}$, whereas $H_{(\lambda=1)}$ represents the Hamiltonian of the final state. The master equation used in TI is[64];

$$\Delta G = \int_{\lambda=0}^{\lambda=1} \langle \frac{\partial H}{\partial \lambda} \rangle_\lambda \, d\lambda \qquad (3.9)$$

where ΔG represents the difference in free energy between state 1 and state 2. The integration of equation 3.9 is evaluated at a series of discrete points between $\lambda = 0$ and $\lambda = 1$. Numerical integration methods are used to calculate ΔG. Selection of enough λ points to make the numerical integration accurate is a primary concern in this method. Therefore, the steps in λ are limited by the requirement to generate enough points to approximate the integral satisfactorily.

The integration of equation 3.9 requires generation of an ensemble average of the derivative of the Hamiltonian with respect to λ at each selected λ point. One can use either MD or MC methods for this purpose. A graphical representation of the TI method is given in Figure 3.1[38]. The area under the curve of $\langle \partial H / \partial \lambda \rangle_\lambda$ vs. λ is the free energy difference between the initial and the final state of a system.

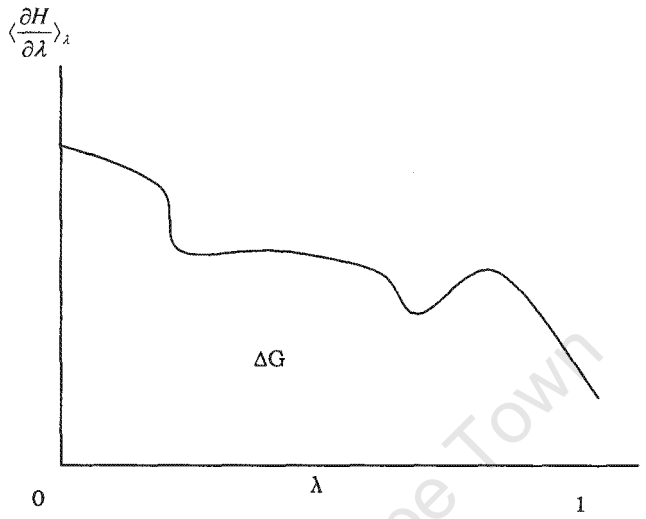


Figure 3.1 Schematic representation of Thermodynamic Integration

3.4 Slow Growth (SG)

In the Slow Growth method, the Hamiltonian of the system is changed only by an infinitesimally small amount[64]. The changes in λ are effected very slowly and therefore the system essentially remains in equilibrium. The master equation used in this method is given by:

$$\Delta G = \sum_{\lambda=0}^{\lambda=1} (H_{(n+1)} - H_{(n)}) \qquad \textbf{(3.10)}$$

It is clear from equation 3.10 that this method does not take an ensemble average into account. In the SG method, the ensemble average is approximated by a single point.

This is done by changing λ incrementally over the full N time steps of a dynamic simulation and using the finite difference between the time steps to approximate the average. Therefore, equation 3.10 can be rewritten as:

$$\Delta G \approx \sum_{i=0}^{N-1} (H_{(\lambda=\lambda_{i+1})} - H_{(\lambda=\lambda i)}) \tag{3.11}$$

In SG , the system is never equilibrated at any λ value and the number of time steps that are required for a given perturbation is not clear[65]. Mainly for these two reasons, this method is not widely used in current practice. However, when the SG methods is used , one should be cautious in applying and interpreting the results [66-68].

3.5 Free Energy Perturbation

In Thermodynamic Integration the change from one state to the other is treated as continuous. The Free Energy Perturbation (FEP) method, first formulated by formulated by Zwanzig[69] , treats the change as a set of discrete steps, called perturbations. According to equation 3.12 the free energy difference *(ΔG)* between states A and B can be expressed as;

$$\Delta G = G(\lambda_B) - G(\lambda_A) = -k_B T \ln \frac{Z(\lambda_B)}{Z(\lambda_A)} \tag{3.12}$$

When an appropriate partition function (NVT partition function is used in this thesis) is substituted in to equation 3.12, it has the form;

$$\Delta G = -k_B T \ln \langle e^{-[H(\lambda_B) - H(\lambda_A)]/k_B T} \rangle_{\lambda A} \tag{3.13}$$

where ΔG is expressed in terms of ensemble averages. The subscript λA indicates that the ensemble average is taken over the ensemble configurations representative of the initial state A. Likewise; it is also possible to give another expression for ΔG with respect to the ensemble average that is representative of state B as:

$$\Delta G = +k_B T \ln \langle e^{-[H(\lambda_A) - H(\lambda_B)]/k_B T} \rangle_{\lambda B} \tag{3.14}$$

This method assumes states A and B to have a good overlap in phase space, as the ensemble average is calculated using configurations of the system that evolve using the potential function of state A, when it goes from A to B. However, the average of the quantity (e.g.: free energy) determined by this method depends on the potential functions of A and B.

Therefore, it is necessary to sample the low energy configurations of both A and B. As the ensemble is generated using the potential function of A, it is relatively less problematic to find the low energy configurations of A. Finding the low energy configuration B will become more difficult if the potential functions of A and B are too dissimilar. If A and B are too dissimilar, the free energy difference calculated will not be accurate as sampling of the phase space of B will not be adequate when simulating A. For this reason, a series of intermediate states are introduced between the original states A and B. These intermediate points are "non-physical" states that connect the physical end points A and B. Then the free energy differences that correspond to each successive set of intermediate points and the overall free energy difference between state A and B is calculated as the sum of above differences, as given by equation 3.15. Since the free energy is a state function, this method still allows one to calculate the free energy differences very accurately. The use of intermediate steps in FEP calculations is depicted in Figure 3.2.

$$\Delta G = G_B - G_A = \sum_{\lambda=\lambda_A}^{\lambda_B} -k_B T \ln\langle e^{-[H(\lambda_B)-H(\lambda_A)]/k_B T} \rangle_{\lambda A} \qquad (3.15)$$

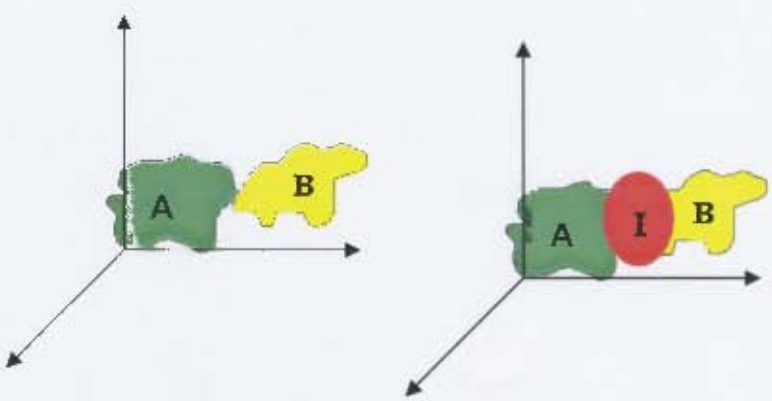


Figure 3.2 Use of intermediates (I) to enhance the overlap between physical end points (A ,B).

3.6 Implementation of Free Energy Perturbation

Free Energy Perturbation is implemented by defining the intermediate points using the coupling parameter (λ) approach. The interval between the physical endpoints A (λ=0) and B (λ=1) is divided into as many intermediate values as desired. These intermediate points can be either equally or unequally spaced. The interval between any successive pair of intermediate points is termed a "window". If there is a total of N number of windows in a simulation, the difference in the free energy between state A and B is given by ,

$$\Delta G = G_B - G_A = \sum_{i=1}^{N} \Delta G_{\lambda(i-1)\rightarrow\lambda(i)} \qquad (3.16)$$

where, $\lambda(i)$, $i=0$, N. *and* i=0 corresponds to state A and i=N corresponds to state B. In an implementation as such, the analytic potential function is made as a function of λ. In a FEP simulation a series of simulations is done by performing a simulation at each λ point and it is done using an appropriately scaled potential function for that particular point. For example, if the CHARMM potential function (equation 3.1) is used in a simulation, at each λ value its Hamiltonian is scaled accordingly. Therefore, in the implementation of FEP, it is essential to re-equilibrate the system to the new potential energy function before a data collection phase.

Because of the above phenomenon, this method has the following general sequence. a). Select λ b). Equilibrate c). Collect data d). Update [λ(i) → λ(i+1)]. In a FEP simulation, the series of simulations when proceeding from initial to final state need not be continuous as each λ point is independent of the other. The ensemble average which is mandatory to solve FEP equations is calculated using MC or MD simulations, where it is done by averaging over the number of moves or the number of integration steps. Calculation of free energy difference by FEP is graphically represented in Figure 3.3, where ΔA represents the free energy difference of the system and λ represents the coupling parameter.

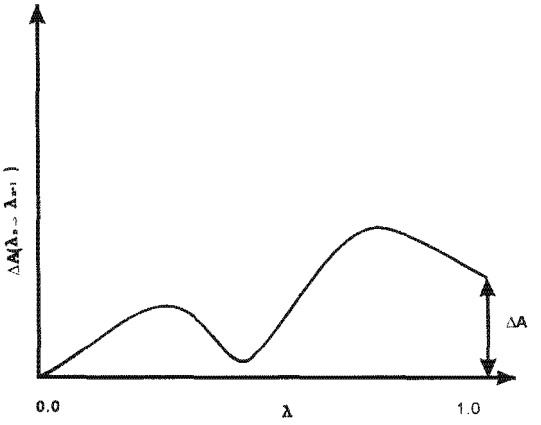


Figure 3.3 Schematic representation of a FEP calculation

3.7 Definition of End Points

In a FEP simulation, the conversion of the system from one state to another must be carried out smoothly, as a function of λ. For this purpose there are three main methods currently available. They are the single topology method[70], dual topology method[71, 72] and hybrid method[73]. Each method has its own strengths and drawbacks and these methods are separately discussed in the following sections.

The single topology method, as the name implies has only one structure, that is a topology exists throughout the perturbation, there being only one set of atoms to represent both starting and end structures (states A and B) of the system. When the system perturbs from states A to B, changes are made to the bonds, valence angles, dihedral angles and non-bonded parameters as λ changes.

In this method dummy atoms are used when the initial number of atoms (state A) does not equal the final number of atoms (State B). A schematic representation of the single topology approach is given in Figure 3.4. There are two significant strengths of this method. Firstly, the convergence may be improved as the changes made to the system at each λ point are smaller. This is because it uses a lower total number of atoms. Secondly, this method allows one to introduce new groups into the system. The main weakness of this method is that it may not be applied to certain perturbations such as going from a closed ring system to an open system or vice versa.

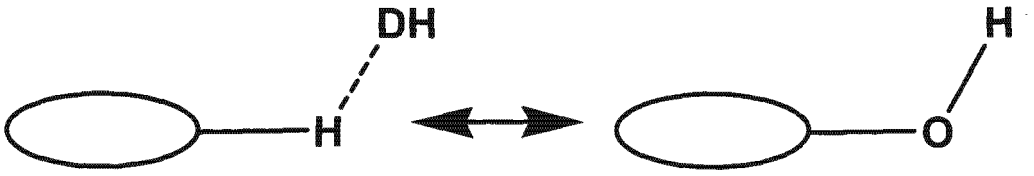


Figure 3.4 Single topology approach. A hydrogen atom is replaced with a hydroxyl group. DH represents the dummy atom, H and OH group do not interact with each other.

In the dual topology method, both topologies (the topology/structure that corresponds to state A and the topology / structure that corresponds to state B) co-exist throughout the perturbation, i.e. there are two groups of atoms involved. The relative weight of the two groups changes as λ changes. However, the atom types and the internal parameters (bond lengths, angles, dihedrals and non-bonded parameters) are never changed in contrast to the single topology method. Furthermore, these two sets of atoms interact with the rest of the system, but never with each other during the simulation. Schematic representation of the dual topology method is given in Figure 3.5.

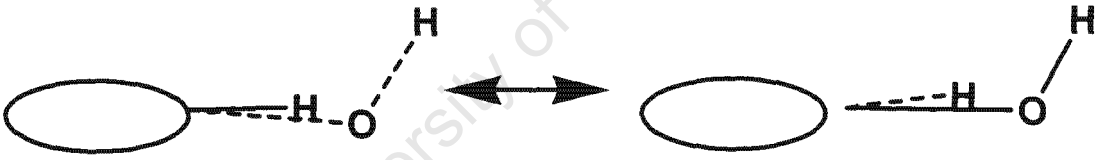


Figure 3.5 Dual topology approach. A hydrogen atom is replaced with a hydroxyl group. Both H and OH co-exist.

The dual topology method is used in this thesis and is the method that is implemented in the CHARMM program. In the dual topology approach the system is divided into four atom groups. They are environmental atoms, co-located atoms, reactant atoms and product atoms. A hybrid Hamiltonian is used to describe the energy of the system. This concept is explained below with the aid of Figure 3.6.

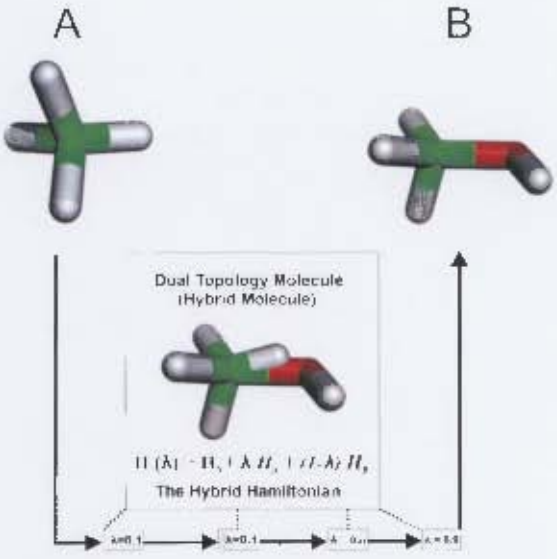


Figure 3.6 Dual topology and hybrid Hamiltonian

The perturbation shown in Figure 3.6 can be carried out if one wants to calculate the free energy difference between molecules A and B. The molecule B differs from molecule A in having a hydroxyl group at the position of a hydrogen atom in A. Therefore in this perturbation a hydrogen atom will be converted in to a hydroxyl group. According to the dual topology method, the hydrogen atom in A and its corresponding hydroxyl group in B are called reactant and product respectively. If the central carbon atom does not change its type but only the charges, it is called a co-located atom and if there are any atoms (other hydrogens) with no change in either type or charge, they are called environment atoms.

The co-existence of both topologies with the equilibrium geometries for both A and B is shown in Figure 3.6. The Hamiltonian that corresponds to any λ point is given by the hybrid Hamiltonian. H_A and H_B, are the Hamiltonians of reactants and products respectively whereas H_0 is the Hamiltonian of the environment and colocated atoms. The major strengths of this method are its applicability to any two end points and the simplicity of the hybrid Hamiltonian in contrast to the Hamiltonian of the single topology method.

3.8 Collecting Data at Different λ points

When the system is changed from one state to another, data are collected at each λ point as explained above. An increase in the number of λ points between two end points will make the calculated free energy difference more accurate. This is achieved by making the perturbation between two successive points smaller. The method that is used in this thesis for data collection is called "double-wide-sampling". The application of this method is schematically represented in Figure 3.7.

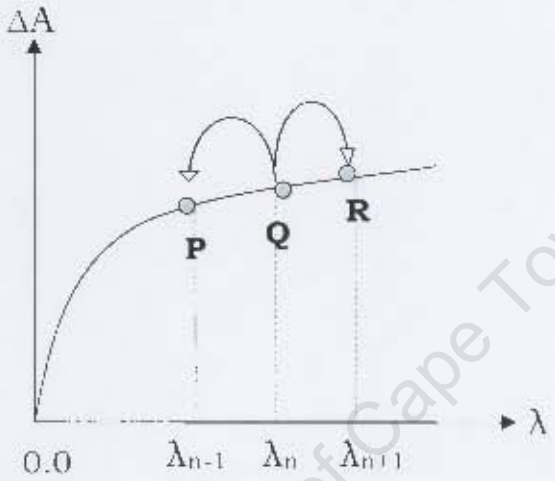


Figure 3.7 Schematic representation of double-wide-sampling method

If Q in Figure 3.7 is an intermediate point (where λ = n) at which a simulation is performed, in the "double wide sampling" method, the same point is used to obtained the free energy differences for $\lambda_n \rightarrow \lambda_{n-1}$ (Q→ P) and $\lambda_n \rightarrow \lambda_{n+1}$ (Q→R).

This method provides a very efficient way to obtain two free energy differences from a single point. Therefore it contributes highly to the accuracy of the overall free energy.

3.9 A General Protocol for Free Energy Calculations

A generalised protocol to perform free energy calculation can be proposed based on the essential steps involved in any type of free energy calculation (TI, SG or FEP). A non–case-specific and method-independent protocol is summarised below.

1. *Defining the starting and the end topologies for the simulation.*

2. *Defining Force Field parameters*

   Correct Force Field parameters must be assigned to both starting and end topologies prior to the simulation. Parameters must be present for both interactions and structure for both end points.

3. *Defining starting coordinates*

   A reasonable (optimised) configuration of the initial state (λ=0) must be used for the simulation.

4. *The system must be energy- minimised, heated and equilibrated prior to the simulation.*

5. *Selection of a free energy method (TI , SG or FEP).*

6 *Selection of the followings for the above selected method*

   Number of λ points, length of equilibration and dynamics to be performed at each λ point.

3.10 Pitfalls in Free Energy Calculations

There are two major sources of error associated with free energy calculations via computer simulations [38] . The first one originates from insufficient sampling of phase space whereas the second originates from inaccuracies associated with Hamiltonians.

The difference in the free energy between two states depends on the sampling of all available phase space. However, it is dominated by the low energy regions along the pathway which connects the end points [74].

Because of this, all free energy calculations have the implicit assumption that the contributions from the regions of the phase space that are not sampled in the simulations to the absolute free energy of the end points cancel. Therefore, the insufficient sampling or the "sampling problem" is an inherited and unavoidable problem in free energy calculations. However, there are few possible changes that one can make to the simulation specifications / protocol in order to improve the results. The simplest of these is to re-run the simulation [74] with sensible changes to the starting coordinates or to the parameters. If the initial velocities are assigned randomly one can assign the velocities with a different random number. These two methods will allow generating a new set of uncorrelated trajectories. Since, the high correlation between successive configurations is one of the major reasons for the "sampling problem" the above mentioned strategy can be used to minimise the errors arising from it.

Performing the equilibration as well as the dynamic phase for longer periods of time generates larger trajectories and allows sampling of more configurations. Thus increasing the duration of equilibration and dynamic phase can be used as a method of improving the final results. The extended length of equilibration will also allow the system to relax properly, whereas a equilibration length shorter than the relaxation time of the system can definitely lead to inaccurate results[75]. An alternative method of improving results in a free energy calculation is to perform the simulation in the reverse direction ($\lambda=1 \rightarrow \lambda=0$) so that the difference in the calculated free energy will give an estimate of the error in the calculation[60].

As mentioned before, the accuracy of the calculated free energy is highly dependant on the descriptors of the interaction in the Hamiltonian. Therefore, force field parameters may lead to large changes in energies and absolute free energies. However, the cancellation effects lead to small changes in relative free energies as a function of the force field [75].

The other common sources of errors in free energy calculations are inappropriate cutoff values and usage of "double wide sampling". The cutoff plays a significant role when there is a creation or annihilation of charges involved in the perturbation[75]. Therefore selecting a correct cutoff value in a FEP simulation is quite significant. The "double wide sampling" has an inherent problem as it uses the same reference ensemble (λ_n) to sample the λ_{n-1} ensemble and λ_{n+1} ensemble and therefore the generated data has been highly correlated[75] . The cost and the accuracy of a free energy calculation are determined by many factors. The nature of the mutation for which the free energy difference is calculated and the size and the flexibility of the system are the dominant factors[38, 60, 75].

3.11 Application of Free Energy Calculations

The application of free energy calculations is two fold. One can devise a method to obtain the absolute free energy (ΔG) for a given change or one can determine the relative free energy differences ($\Delta\Delta G$) for given processes. Even though the first is applicable to certain cases, because of its high inaccuracy the majority of the applications are found in the realm of the second category. Sections 3.11.1 and 3.11.2 describe some practical implementation of these methods. Each method uses an appropriate thermodynamic cycle to calculate the desired free energy value. This is generally referred to as the "thermodynamic cycle approach".

3.11.1 Calculation of Absolute Free Energies

There are two major thermodynamic cycles that can be used to calculate the absolute free energies of association, one proposed by Cieplak and Kollman [76] and the other by Jorgensen *et al*[77].

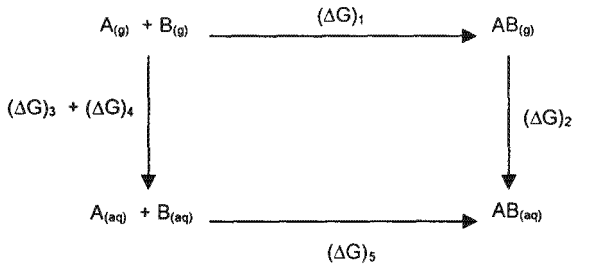


Figure 3.8A free energy cycle to calculate Absolute free energy of association [76]

Figure 3.8 shows the scheme proposed by Cieplak and Kollman where, ΔG_1 is the gas-phase free energy of association, which is calculated using molecular mechanics energy minimisation and normal mode analysis[64]. The free energies for solvation of A (ΔG_3), B (ΔG_4) and AB (ΔG_5) can be calculated by mutating each of these species to nothing in the solution, and the absolute free energy of binding can be calculated according to Equation 3.17.

$$\Delta G_5 = \Delta G_1 + \Delta G_2 - \Delta G_3 + \Delta G_4 \qquad \textbf{(3.17)}$$

The scheme proposed by Jorgensen *et al.* is shown in Figure 3.9.

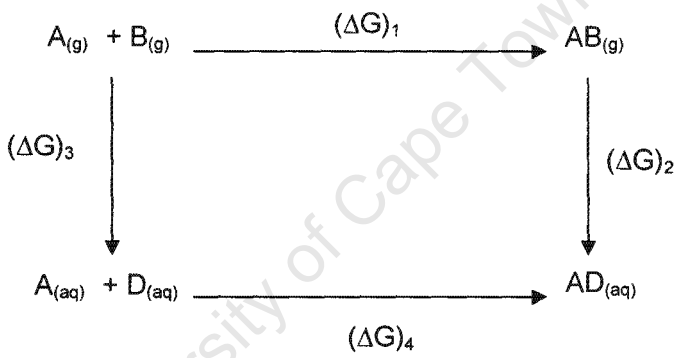


Figure 3.9: A free energy cycle to calculate Absolute free energy of association[77].

In Jorgensen's method (Figure 3.9) a dummy molecule is used to simplify the calculation where, A and B are real molecules whereas D is a dummy molecule. This method requires only 2 mutations to be done, that is B to nothing in water and B to nothing in the AB complex. The free energy of association is given by:

$$\Delta G_1 = \Delta G_2 - \Delta G_3 \qquad \textbf{(3.18)}$$

3.11.2 Calculation of Free Energy Differences

There are two major categories of molecular phenomena that are studied using free energy difference calculations: (1) Solvation and (2) Molecular association [60]. The relative solvation free energy of two solutes (A and B) can be calculated by using the thermodynamic cycle given in Figure 3.10.

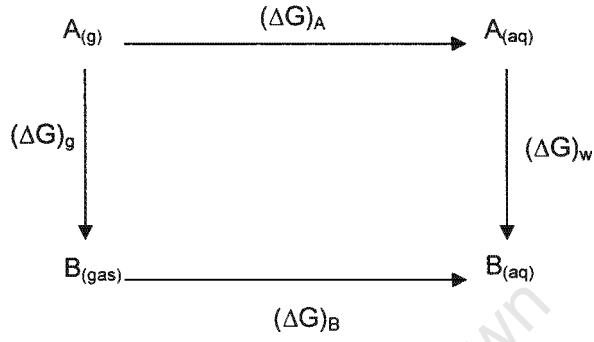


Figure 3.10 A free energy cycle to calculate relative solvation free energy

Application of the thermodynamic cycle given in Figure 3.10 requires two mutations to be done. Solute A is mutated to solute B in the gas phase and the same mutation is repeated in water. The relative solvation free energy ($\Delta\Delta G_{sol}$) is given by Equation 3.19.

$$\Delta\Delta G_{sol} = \Delta G_B - \Delta G_A = \Delta G_w - \Delta G_g \qquad \textbf{(3.19)}$$

The free energy calculations on molecular associations can be obtained by using thermodynamic cycles like the one given in the Figure 3.11.

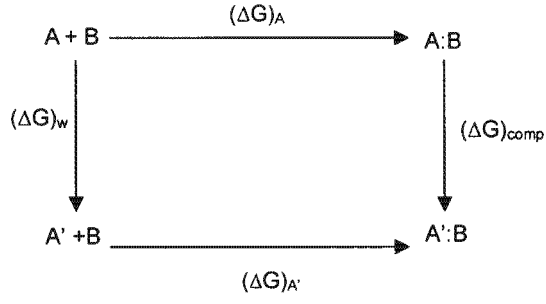


Figure 3.11 : A free energy cycle to calculate relative association free energy

If A and A' are two different molecules, the relative free energy for their association with B can be calculated using the above thermodynamic cycle. A is mutated to A' in water (for ΔG_w) and the same mutation is repeated in complex with B in water (for ΔG_{com}).

The relative free energy of association is given by Equation 3.20.

$$\Delta\Delta G_{assc} = \Delta G_{A'} - \Delta G_A = \Delta G_{comp} - \Delta G_w \quad \textbf{(3.20)}$$

If A and A' are two ligands and B is a protein, the above calculated $\Delta\Delta G_{assc}$ will be the relative free energy of binding for these ligands.

3.12 Potential of Mean Force (PMF)

In the previous sections it was described how to perform free energy calculation as a function of chemical mutation (structural change). If the parameter linking the initial and the final state is a function of special coordinates of the system, the calculated free energy difference is referred to as the potential of mean force (PMF). The simplest type of PMF is the free energy change as a function of the distance between two particles. The PMF calculations however suffer from inadequate sampling of regions where conformations are drastically different from the 'most likely' conformation and this is generally referred to as not sampling the high energy conformations. To avoid this problem, the technique of umbrella sampling is traditionally used.

3.13 Why Difference in Free Energy?

The difference in free energy is expressed only in terms of perturbed interactions but not in terms of the system as a whole[74]. The change in internal energy / enthalpy or entropy is expressed in terms of a whole system. Therefore the error in calculated enthalpy / entropy changes would be very large compared to the error in calculated free energy. Even though there are efficient ways of calculating enthalpy and entropy using FEP and TI, free energy calculations are widely used, because of their high accuracy.

CHAPTER 4

Rationalising the Effect of Selected Inhibitors

4.1 Introduction:

Ricin is a heterodimeric (A-B type) toxin isolated from *Ricinus communis*[10]. Ricin contains a total of 516 amino acids[78] with the toxic A chain (RTA) consisting of 267 amino acids. Recently, there have been numerous applications found for ricin because of its high toxicity. A hypothetical binding pocket has been proposed for RTA by various studies [10, 79, 80]. As previously stated (Section 1.7) some key amino acid residues have been identified. TYR 80, VAL 81, TYR 123, GLU 177 and ARG 180 are often reported in literature as key residues in RTA catalytic mechanism although their exact roles in the mechanism are not yet clear. The cytotoxic activity, proposed mechanism and the binding site of ricin are discussed in section 1.7 in detail. The structure of Ricin and its proposed binding site are presented in Figure 4.1.

Figure 4.1 Structure of RTA and its key residues

4.2 Ricin Inhibitors

There is no effective inhibitor for ricin. However, numerous studies have attempted to identify ricin inhibitors using experimental [10, 19, 80] and/or computational[81-83] methods. The inhibitors reported in literature can be classified into three distinct classes as monocyclic, bicyclic or tricyclic, depending on the number of rings in their structures. The most frequently reported RTA inhibitors are summarised in Table 4.1. The monocyclic inhibitors show minimal or no inhibition activity[10]. Three-ring inhibitors (tricyclic) have shown sufficient inhibitor activity against RTA[10, 25]. The pterin-like tricyclic inhibitors were discontinued owing to their poor solubilities and adenosine like three-ring inhibitors (e.g. formycin monophosphate) were identified as non-powerful RTA inhibitors[25]. Inhibitors with heterobicyclic structures which resemble adenine base (a product of the RTA catalytic mechanism) have shown a considerable degree of success as RTA inhibitors[10]. In this study it was attempted to rationalise the effect of guanine-like ricin inhibitors (i.e. inhibitors that resemble adenine, the product in core structure) with the aid of computational methods. Inhibitors that were selected for this study [10] are presented in Table 4.2.

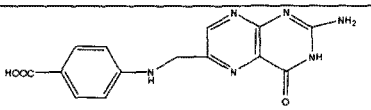
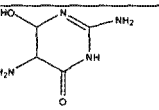
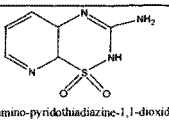
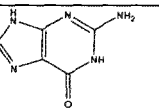
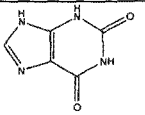
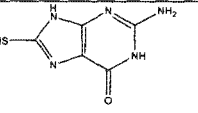
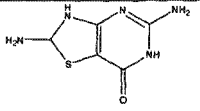
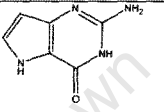
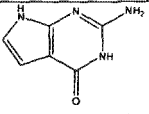
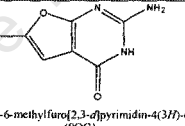
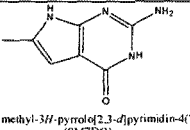
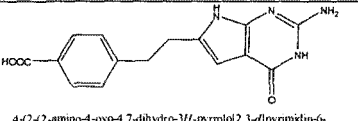
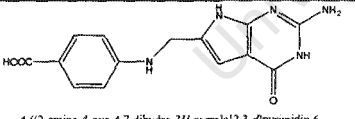
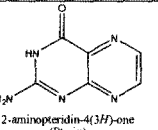
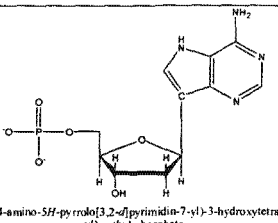
2D Structure and Name	IC_{50} mM	2D Structure and Name	IC_{50} mM
 4-((2-amino-4-oxo-3,4-dihydropteridin-6-yl)methylamino)benzoic acid (Pteroic acid)	0.6	 2,5-diamino-6-hydroxy-5,6-dihydropyrimidin-4(3*H*)-one (DDP)	2.2
 3-amino-pyridothiadiazine-1,1-dioxide	No inibitio n	 2-amino-1*H*-purin-6(9*H*)-one (Guanine)	0.9
 1*H*-purine-2,6(3*H*,9*H*)-dione (Xanthine)	3.6	 2-amino-8-mercapto-1*H*-purin-6(9*H*)-one	0.56
 2,5-diamino-2,3-dihydrothiazolo[4,5-*d*]pyrimidin-7(6*H*)-one	2.0	 2-amino-3*H*-pyrrolo[3,2-*d*]pyrimidin-4(5*H*)-one (9DG)	1.4
 2-amino-3*H*-pyrrolo[2,3-*d*]pyrimidin-4(7*H*)-one (7DG)	2.8	 2-amino-6-methylfuro[2,3-*d*]pyrimidin-4(3*H*)-one (9OG)	0.4
 2-amino-6-methyl-3*H*-pyrrolo[2,3-*d*]pyrimidin-4(7*H*)-one (8M7DG)	2.1	 4-(2-(2-amino-4-oxo-4,7-dihydro-3*H*-pyrrolo[2,3-*d*]pyrimidin-6-yl)ethyl)benzoic acid	0.6
 4-((2-amino-4-oxo-4,7-dihydro-3*H*-pyrrolo[2,3-*d*]pyrimidin-6-yl)methylamino)benzoic acid	0.7	 2-aminopteridin-4(3*H*)-one (Pterin)	-
 ((2*R*,3*R*,5*R*)-5-(4-amino-5*H*-pyrrolo[3,2-*d*]pyrimidin-7-yl)-3-hydroxytetrahydrofuran-2-yl)methyl phosphate (Formycin-5'- Monophosphate)	2.5		

Table 4.1 Ricin Inhibitors

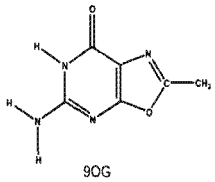
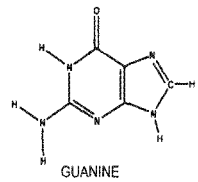
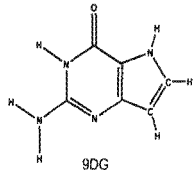
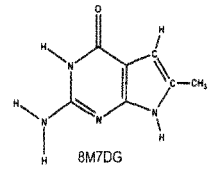
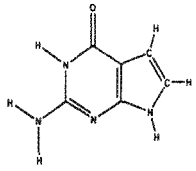
Structure	IC_{50} (mM)	Structure	IC_{50} (mM)
 9OG	0.4	 GUANINE	0.9
 9DG	1.4	 8M7DG	2.1
 7DG	2.8		

Table 4.2 Selected Inhibitors and their IC_{50} Values

(Extracted from table 4.2)

All inhibitors shown in Table 4.1 have guanine as their structure template and they all have a heterobicyclic (5-membered and 6-membered rings) moiety in common with adenine (product). Among these inhibitors, guanine is structurally significant as the rest of the inhibitors are derivatives of it. The 6-membered moieties of all inhibitors are identical and the inhibitors can be differentiated from one another according to the changes in the 5-membered ring. According to the experimental IC_{50} values, the inhibition power of the above inhibitors are in the order of 9OG > GUANINE > 9DG > 8M7DG >7DG, where 9OG is the best inhibitor and 7DG is the weakest.

Structure-Based Drug Design (SBDD) is one of several methods used in the drug discovery process [84]. In SBDD the three-dimensional structure of the target molecule (enzyme, receptor etc) is used to assist the development of new drug compounds. The structure of the target molecule is usually obtained from x-ray crystallographic and / or NMR experiments.

The most important feature of this technique is its ability to reveal a high level of details about the interactions between ligands and the target molecule[84]. The structure-based approach becomes even more powerful when the structures of both target molecule and at least one of its active ligands are known. In such a situation one can use this approach to propose new ligands with maximised target-ligand interactions i.e. to propose new drug compounds. In this study it was attempted to answer the question; are "product-like" (guanine-based) inhibitors powerful? If the answer to this question can be obtained, it can be used to predict the structural and chemical features of a better inhibitor. To seek an answer to this question a structure-based approach was employed in this thesis.

Firstly, it was attempted in this study to calculate the relative free energies of binding for the selected inhibitors using Free Energy Perturbation (FEP) methods (methodology described in Section 3.5). From this one will be able to rank the ligands in the order of increasing inhibition. Secondly, molecular dynamics methods were used to completely understand the RTA binding pocket, the dynamics behaviour of ligand-RTA complexes and the mode of binding of different ligands (inhibitors, substrate and product) to RTA.

4.3 Methodology

4.3.1 Molecular Dynamics Simulations

In the present study of ricin, molecular dynamics simulations were performed for all ricin:inhibitor complexes as well as for the ricin:substrate and ricin:product complexes. For all molecular dynamics simulations the CHARMM33b2[41] programme was used. An empirical energy function that contains terms for both internal and external interactions is used in CHARMM programme this is shown in Equation 4.1.

$$V(\vec{R}) = \sum_{bonds} k_d (d - d_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} k_\chi [1 + \cos(n\chi - \delta)]$$
$$+ \sum_{impropers} k_\varphi (\varphi - \varphi_0)^2 + \sum_{Urey-Bradley} k_{UB} (S - S_0)^2$$
$$+ \sum_{non-bonded} \{\varepsilon_{ij} [(R_{ij}^{\min} / r_{ij})^{12} - (R_{ij}^{\min} / r_{ij})^6] + q_i q_j / e_l r_{ij}\}$$

(4.1)

The protein was modelled using CHARMM27[85, 86] all atom force field for protein and nucleic acids. Initial preparation of the protein and the parameterisation of the inhibitors are discussed in section 4.3.2. In all molecular dynamic simulations a water sphere of 23.4 Å containing 1941 TIP3P water molecules was used. The system was centred on the C5 (bridge carbon) carbon of the ligand for the simulations. Those water molecules that overlapped with any of the ligand heavy atoms were removed and a spherical boundary force was applied to the surface of the water sphere in order to maintain a consistent water density of 1.0 g cm^{-1}. A Constant temperature of 298.15 K was maintained in simulations. A buffer region of 2Å thickness and a dynamic region with 21Å radius were used with Langevin dynamics in all simulations. All hydrogen bond lengths were kept fixed with the SHAKE [87] algorithm and the integration step of 1fs was used. The long-range interactions were truncated using a switch function and for that the selection of atoms was done in group-by-group fashion.

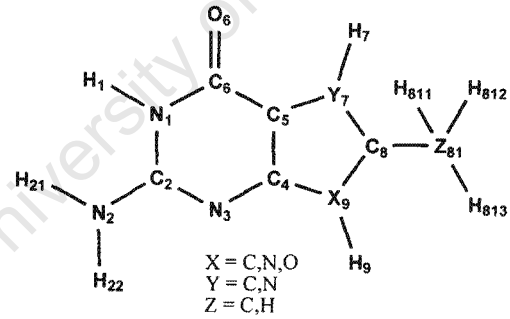
4.3.2 Initial Preparation and Parameterisation

The inhibitors were modelled with newly introduced parameters to the Equation 4.1. Parameterisation of bond, angle, dihedral and non-bonded terms was done by analogy where as partial atomic charges were obtained from quantum mechanical calculations. The partial atomic charges were calculated using the Merz-Kollman charge fitting method[88] at B3LYP[59]/6-31G* level of theory. Charges were scaled in order to fit to the force field by using guanine as a reference. The atom types used for the inhibitors and their partial atomic charges are presented in Table 4.3. Parameterisation is discussed in more details in Section 2.13.

Inhibitor												
9OG			9DG			8M7DG			7DG			
Atom	Atom Type	Atomic Charge	Atom	Atom Type	Atomic Charge	Atom	Atom Type	Atomic Charge	Atom	Atom Type	Atomic Charge	
O9	ON6B	-0.20	C9	CA	-0.11	N9	NN2B	-0.11	C7	CA	-0.41	
C4	CN5	0.50	C4	CN5	0.25	H9	HN2	0.32	N9	NN2B	-0.11	
N2	NN1	-0.96	H9	HP	0.12	C4	CN5	0.42	C4	CN5	0.27	
H21	HN1	0.41	H7	HN2	0.33	N2	NN1	-0.72	H9	HN2	0.32	
H22	HN1	0.43	N2	NN1	-0.89	H21	HN1	0.31	H7	HP	0.16	
N3	NN3G	-0.75	H21	HN1	0.39	H22	HN1	0.35	N2	NN1	-0.72	
C2	CN2	0.86	H22	HN1	0.39	N3	NN3G	-0.79	H21	HN1	0.32	
N1	NN2G	-0.56	N3	NN3G	-0.77	C2	CN2	0.88	H22	HN1	0.35	
H1	HN2	0.34	C2	CN2	0.86	N1	NN2G	-0.41	N3	NN3G	-0.79	
C6	CN1	0.67	N1	NN2G	-0.29	H1	HN2	0.26	C2	CN2	0.85	
O6	ON1	-0.63	H1	HN2	0.22	C6	CN1	0.63	N1	NN2G	-0.41	
C5	CN5G	0.01	C6	CN1	0.61	O6	ON1	-0.54	H1	HN2	0.26	
N7	NN2B	-0.56	O6	ON1	-0.56	C5	CN5G	-0.25	C6	CN1	0.63	
C8	CN1	0.50	C5	CN5G	-0.27	C7	CA	-0.41	O6	ON1	-0.54	
C81	CT3	-0.48	N7	NN2B	-0.41	H7	HP	0.18	C5	CN5G	-0.25	
H811	HA	0.12	C8	CA	-0.07	C8	CN1	-0.13	C8	CA	-0.13	
H812	HA	0.15	H8	HP	0.20	C81	CT3	-0.27	H8	HP	0.20	
H813	HA	0.15				H811	HA	0.09				
						H812	HA	0.09				
						H813	HA	0.09				

Table 4.3 Atom types and their partial atomic charges for newly parameterised ligands

(The key given below is for identification of the positions given in the Table 4.2)



Key for Table 4.3.

The molecular simulations were initiated with the crystal structures obtained from the Brookhaven Protein Data Bank (1IL3, 1IL4, 1IL9, 1FMP). The crystals structures have been resolved at 2.8Å , 2.6Å , 3.0Å and 2.8Å for 1IL3, 1IL4, 1IL9[10] and 1FMP[25] respectively. All protein structures were checked with WHATIF program[57] for quality of the crystal structure. 1IL3 was found as the best structure and therefore it was used for the rest of the ricin:inhibitor simulations.

In the situations of guanine, and 8M7DG where no crystal structures were available, 9OG and 7DG coordinates were used as starting coordinates for guanine and 8M7DG respectively.

The protonation states of the titratable groups were determined by using the pKa tool of the WHATIF[57] program. All simulations were performed at constant pH of 7.00 and therefore the protonated states were determined at this pH. The protein was assigned an overall charge of +2.00 e due to the fact that amino acids HIS 106 and HIS 40 were in the protonated state. The determination of protonation states and initial preparation of a protein for a computer simulation in discussed in detail in Sections 2.12.1 and 2.12.2.

4.3.3 Free Energy Perturbation Calculations

Free Energy Perturbation (FEP) method can be used to calculate the relative free energy differences for a given set of processes. FEP is a statistical mechanical approach which allows calculating the thermodynamic properties of the system of interest from its microscopic descriptors. In the FEP method the change (of the system of interest) from one state to the other is treated as a set of discrete steps. These intermediate steps in the conversion pathway are linked through a coupling parameter λ. The free energy difference (ΔG) between states A and B is given by equation 4.2 [69].

$$\Delta G = G(\lambda_B) - G(\lambda_A) = -k_B T \ln \frac{Z(\lambda_B)}{Z(\lambda_A)} \qquad (4.2)$$

Equation 4.2 can be re-written in the form of equation 4.3 with partition (Z) for NVT ensemble which is used in our simulations.

$$\Delta G = -k_B T \ln \langle e^{-[H(\lambda_B) - H(\lambda_A)]/k_B T} \rangle_{\lambda A} \qquad (4.3)$$

In equation 4.3 $\langle\ \rangle$ represents the ensemble average taken over the configurations representative of state A.

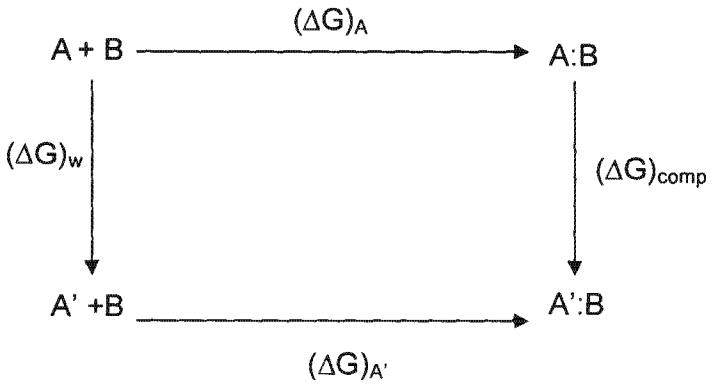


Figure 4.2 Thermodynamic cycle for calculating ΔΔG

One can calculate the relative free energy of binding of two inhibitors using a thermodynamic cycle similar to the one shown in Figure 4.2. The free energy of binding for inhibitor A and A' are given by (ΔG)A and (ΔG)A' respectively whereas (ΔG)w represents the difference in solvation free energy between A and A'. The difference in free energy of binding of A and A' to the active site of the protein is given by $(\Delta G)_{comp}$. Simulations were performed to calculate (ΔG)w and $(\Delta G)_{comp}$. Inhibitor A was perturbed to A' in water with molecular dynamics and the same mutation was repeated in the active site of the protein in the presence of water . The relative free energy of binding was then obtained by using Equation (4.4).

$$\Delta\Delta G_{binding} = \Delta G_{A'} - \Delta G_{A} = \Delta G_{comp} - \Delta G_{w} \quad \textbf{(4.4)}$$

Four free energy perturbations (9OG→Guanine, Guanine→ 9DG , 9DG→ 8M7DG and 8M7DG→ 7DG) were performed. These perturbations were specifically chosen to represent the IC_{50}-based ranking of those inhibitors. In each perturbation, a stronger inhibitor (low IC_{50}) was mutated in to a weaker (high IC_{50}) inhibitor. The thermodynamic cycle given in Figure 4.2 was used to calculate the relative free energy of binding (ΔΔG) of these inhibitors with ricin where a +ΔΔG value would indicate a stronger binding of the first inhibitor with respect to the second one. All FEP simulations were performed using CHARMM program with molecular dynamics.

The same simulation conditions as mentioned in the previous section were used for all FEP simulations. The dual topology method was used with windows of λ= 0.05. At the end point the largest change in the system occurs: a non-interacting / partially-interacting atom changes to an interacting one or vice versa. As a measure to minimise this end-point problem i.e. to improve the convergence at the end points and to make the perturbation smoother, a λ value of 0.025 was used at the foremost and 0.975 at the final most windows. The starting coordinates for the hybrid were obtained from the starting inhibitor's coordinates in the perturbation after 3ns of equilibration in the binding site. At each window, 500ps of equilibration was performed followed by 500ps of data collection. All FEP results were post-processed by using CHARMM program (TSM functionality with double wide sampling technique).

4.4 Results and Discussion

4.4.1 Free Energy Perturbation Results

All inhibitors were ranked using the relative free energies of binding ($\Delta\Delta G_{binding}$) which were obtained from free energy perturbation simulations. These results are summarised in Table 4.4. The ranking of the inhibitors obtained from the computational results is in perfect agreement with the IC_{50} -based ranking and it is in the order of 9OG > GUANINE > 9DG > 8M7DG >7DG. These results therefore, reconfirm the experimental data and more importantly, prove the validity of the force field parameters and computational techniques used.

Perturbation	ΔΔG (kcal/mol)
9OG(0.4) → Guanine(0.9)	1.37
Guanine(0.9) → 9DG(1.4)	0.45
9DG(1.4) → 8M7DG(2.1)	1.03
8M7DG(2.1) → 7DG(2.8)	1.48

Table 4.4 Relative binding free energies of the selected inhibitors

The IC_{50} values are given in the brackets.

4.4.2 Nature of Inhibitor Binding

Trajectories of 15 ns molecular dynamics simulations were analysed to investigate the nature of binding of each inhibitor, substrate (adenosine) and the product (adenine) to RTA. Analysis was done in the mean of non-bonded (electrostatic and Van der Waals) interactions, hydrogen bonding, root mean square deviation (RMSD) of the structure and some distance based- criteria.

The previous studies[10] of these inhibitors suggest that hydrogen bonding (H-bonding) and pi stacking play significant roles in binding of these inhibitors. The trajectories from MD simulations were analysed to check for hydrogen bonding. A hydrogen bond was defined to be present only if the donor-hydrogen-acceptor angle is between 180° - 120° and the hydrogen-acceptor distance is less than 3Å. Hydrogen bonding in the binding of each ligand is discussed later.

An aromatic interaction (pi-stacking / π-π interaction) is a noncovalent interaction between the aromatic moieties of organic molecules. The overlapping of p-orbitals of conjugated systems in close proximity gives rise to pi-stacking[89]. The relative spatial orientation of participating systems is therefore crucial in this scenario. Among many measurable parameters, the distance between the participating pi systems is a straightforward criterion and can be easily measured. This distance is usually reported to be 3.00 – 4.00 Å for effective pi stacking [90].

The distance between the centre of geometry (COG) of the six-membered ring of the ligand (for all ligands) and the COG of the phenyl ring of the residues TYR 80 and TYR 123 were measured using the MD trajectories. These results are presented in Table 4.5. According to these results, none of the ligands are capable of making pi interactions with TYR 123 as the distance criteria fails. Therefore, it is highly unlikely that these ligands will orient in such a manner that they will be "sandwiched" between TYR 80 and TYR 123 as suggested by some previous crystallographic results[10]. Much shorter distances between TYR 80 and the inhibitors 9OG, Guanine and the substrate adenosine might be an indication of a pi-pi interaction between them.

However, since all these distances are greater than 4.00Å it may be assumed that if pi-pi interactions do indeed exist, they would be extremely weak and would minimally affect the overall binding energy.

Ligand	Mean Distance (Å)	
	TYR 80	TYR 123
9OG	4.08	5.36
Guanine	4.15	7.52
9DG	7.31	7.03
8M7DG	4.9	5.49
7DG	7.30	8.69
Adenosine (substrate)	3.8	11.25
Adenine base (product)	5.62	5.02

Table 4.5 Mean distance between COG of the six membered rings of the ligands and COG of the phenyl ring of TYR 80 and TYR 123

The H-bond analysis and pi interaction analysis (distance search) clearly suggest that these inhibitors predominantly use electrostatic and / or van der Waals (vdW) interactions to bond to RTA. The non-bonded interactions between ligands and RTA were calculated using the trajectories. These results are summarised in Table 4.6. (Only the total *average* interaction energies that are greater than -1 kcal/mol are included in this table). The relative orientation of the interacting amino acids and the ligands are schematically presented in the Figure 4.3- Figure 4.9.

	Residue	Ligand / Average Interaction Energy (kcal/mol)																				
		9OG (1)			Guanine (2)			9DG (3)			8M7DG (4)			7DG (5)			Adinine			Adenine(Base)		
		Total	VWD	ELEC	Total	VWD	ELEC	Total	VWD	ELEC	Total	VWD	ELEC	Total	VWD	ELEC	Total	VWD	ELEC	Total	VWD	ELEC
Region A	TYR 80	-7.50	-6.76	-0.73	-7.36	-5.55	-1.81	-1.12	-1.14	0.01	-4.33	-3.93	-0.40	-2.65	-2.38	-0.28	-8.75	-6.45	-2.30	-3.32	-2.43	-0.90
	VAL 81	-9.80	-1.01	-8.79	-10.83	-1.46	-9.37							-2.36	-0.45	-1.91	-6.92	-1.21	-5.71			
	GLY 121	-1.40	-0.19	-1.20							-3.65	0.24	-3.89	-1.27	-0.61	-0.65	-2.09	-0.96	-1.13	-2.30	-0.47	-1.83
	ASN 122	-1.42	-0.36	-1.06				-2.49	-1.34	-1.15	-2.66	-0.82	-1.84	-2.61	-0.84	-1.76				-4.17	-1.45	-2.72
	TYR 123	-5.96	-3.01	-2.95	-2.67	-0.72	-1.96	-2.06	-1.40	-0.66	-3.45	-2.37	-1.09	-2.00	-0.91	-1.10				-5.14	-4.12	-1.02
	ILE 172	-2.76	-2.89	0.14	-5.67	-3.52	-2.14							-1.36	-1.23	-0.13	-4.43	-4.57	0.14			
	ARG 134	-3.07	-0.12	-2.95	-5.74	-0.14	-5.60							-1.04	-0.04	-1.00						
	ASP 100				-1.06	-0.11	-0.94	-7.38	-0.16	-7.23				-1.83	-0.06	-1.76						
	Total A	-31.90	-14.34	-17.55	-33.32	-11.50	-21.82	-13.06	-4.03	-9.03	-14.09	-6.88	-7.21	-15.11	-6.52	-8.60	-22.19	-13.19	-9.00	-14.93	-8.47	-6.47
Region B	SER 176																			-1.16	-0.33	-0.83
	GLU 177																					
	ARG 180																-7.04	-1.06	-5.98	-8.27	-0.24	-8.03
	GLU 208																-17.50	0.53	-18.02			
	TRP 211																-2.44	-1.75	-0.69			
	Total B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-26.98	-2.28	-24.70	-9.43	-0.57	-8.86
Non-specific amio acids (X)	ASP 75										-1.98	-0.24	-1.74									
	ASN 78										-1.66	-1.20	-0.46									
	PHE 93				-1.69	-1.54	-0.15															
	PRO 95										-2.17	-1.14	-1.03									
	ASN 97																					
	ASP 124																			-3.58	-0.30	-3.28
	ARG 125													-1.59	-0.31	-1.27						
	PHE 168																-2.32	-2.00	-0.33			
	Total X	0.00	0.00	0.00	-1.69	-1.54	-0.15	0.00	0.00	0.00	-5.82	-2.58	-3.23	-1.59	-0.31	-1.27	-2.32	-2.00	-0.33	-3.58	-0.30	-3.28
	Total ABX	-31.90	-14.34	-17.55	-35.01	-13.04	-21.97	-13.06	-4.03	-9.03	-19.91	-9.47	-10.44	-16.70	-6.83	-9.87	-51.50	-17.47	-34.03	-27.95	-9.35	-18.61
	Ricin	-36.78	-19.86	-16.92	-41.11	-18.14	-22.97	-24.25	-9.12	-15.13	-28.97	-16.91	-12.06	-27.14	-13.28	-13.86	-57.46	-25.98	-31.48	-32.30	-14.57	-17.74

Table 4.6. Average interaction energies between ligands and key amino acid residues

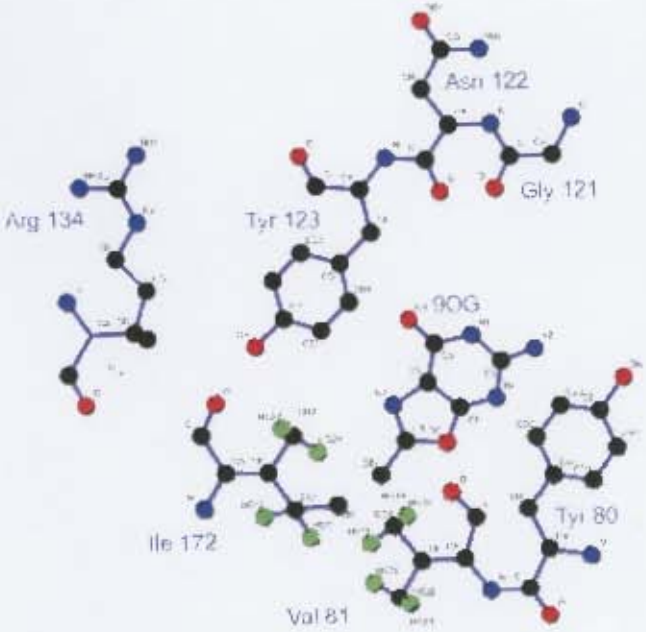


Figure 4.3 90G - Region A contacts

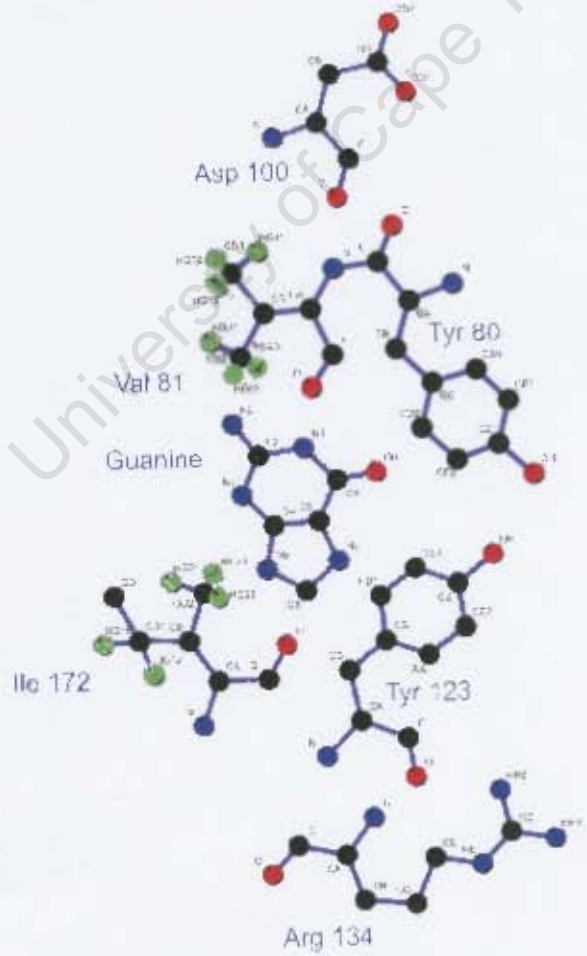


Figure 4.4 Guanine - Region A contacts

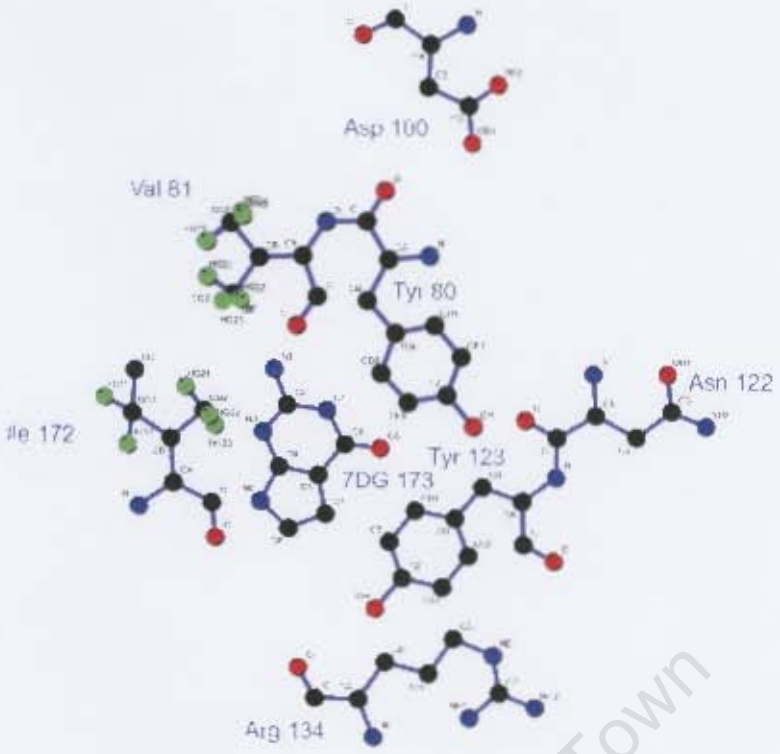


Figure 4.7 7DG - Region A contacts

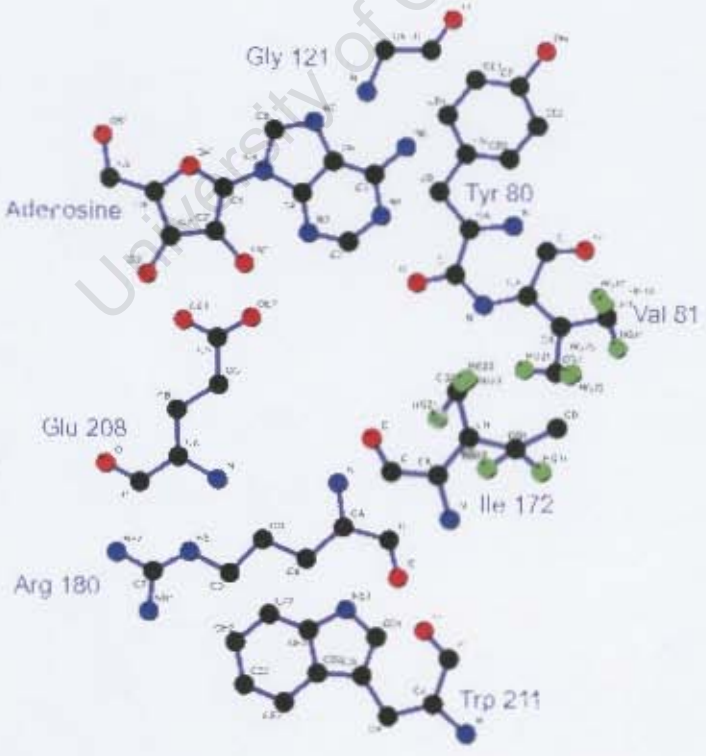


Figure 4.8 Adenosine - Region A contacts

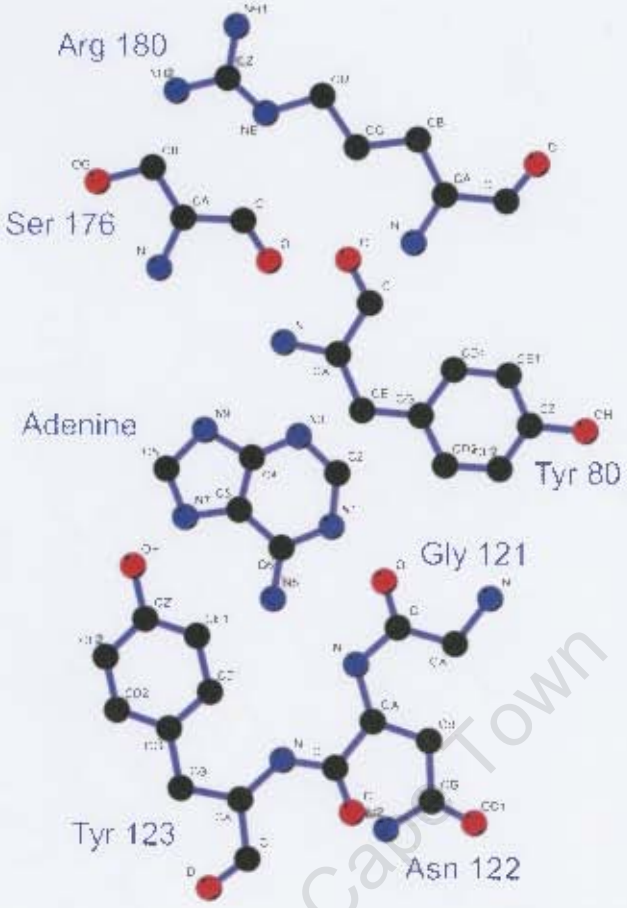


Figure 4.9 Adenine - Region A contacts

Based on the degree of RTA-ligand interactions, two distinct areas of RTA binding pocket were clearly identified. The first region consists of the residues TYR 80 ,VAL 81, GLY 121 , ASN 122, TYR 123 , ILE 172 , ARG 134 , ASP 100. The second region has the residues SER 176, GLU 177, ARG 180, GLU 208 and TRP 211. From this point onwards in this discussion the first region will be referred to as region A and the second region as region B. These regions are schematically presented in the Figure 4.10 and they are highlighted in the RTA structure in Figure 4.11

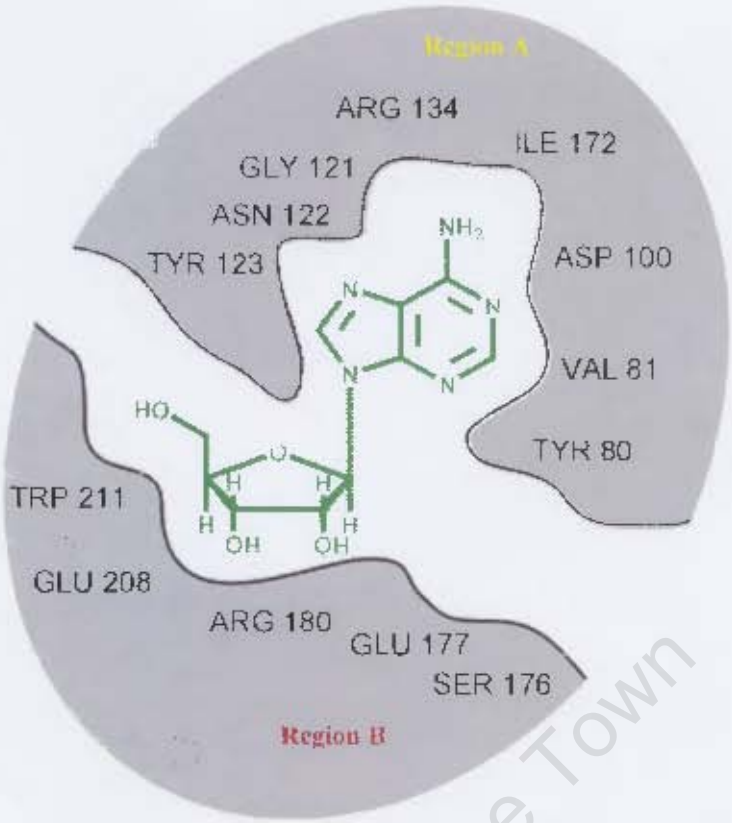


Figure 4.10 Schematic representation of region A and B of RTA

Figure 4.11 3D-structure of Region A (yellow) and Region B(red)of RTA binding site.

The interaction energies given in Table 4.6 imply that region A and B are significant in binding of those inhibitors to RTA. For a example, 90% of substrate's (adenosine) interaction energy originates from the interactions with region A and B and all inhibitors bind to RTA via the interactions with region A. The binding of each ligand to RTA is described individually in the following sections (key to Table 4.3 on page 72 is used to refer to atom numbers).

4.4.2.1 Binding of 9OG

9OG is the best among the selected inhibitors according to the experimental IC_{50} values. The free energy perturbation results of this study place 9OG at the same position in the inhibition hierarchy. The significant structural feature of this inhibitor with respect to the other inhibitors is the presence of oxygen at the 9^{th} position. Binding of 9OG occurs mainly via electrostatic and van der Waals interactions and it is facilitated by two key H-bonding sites. Both N3 and O9 receive H-bonding from amine hydrogen of VAL 81. However, the hydrogen bond between O9 and amine hydrogen of VAL 81 is intermittent in its nature whereas the N3-VAL 81(amine hydrogen) hydrogen bond is continuous.

The binding of 9OG occurs only via the residues in region A and it does not interact with any of the residues in region B. The residues TYR 80, VAL 81, and TYR 123 are significantly involved in the binding of this inhibitor. The participation of ARG 134 in binding of 9OG is very remarkable as ARG 134 has not been reported previously as an active amino acid in RTA. The total interaction energy of 9OG with region A is -31.90 kcal/mol. Over 50% of non-bonded interaction energy is electrostatic and the rest originates from van der Waals interactions.

It can be clearly seen from Table 4.6 that the interactions with residues TYR 80, TYR 123 and ILE 172 are predominantly vdW while the interactions with VAL 81 and ARG 134 are mainly electrostatic. The interaction energy time series of 9OG with region A is presented in Figure 4.12.A. It shows that 9OG-region A interaction is continuous and steady throughout the simulation period of 15 ns.

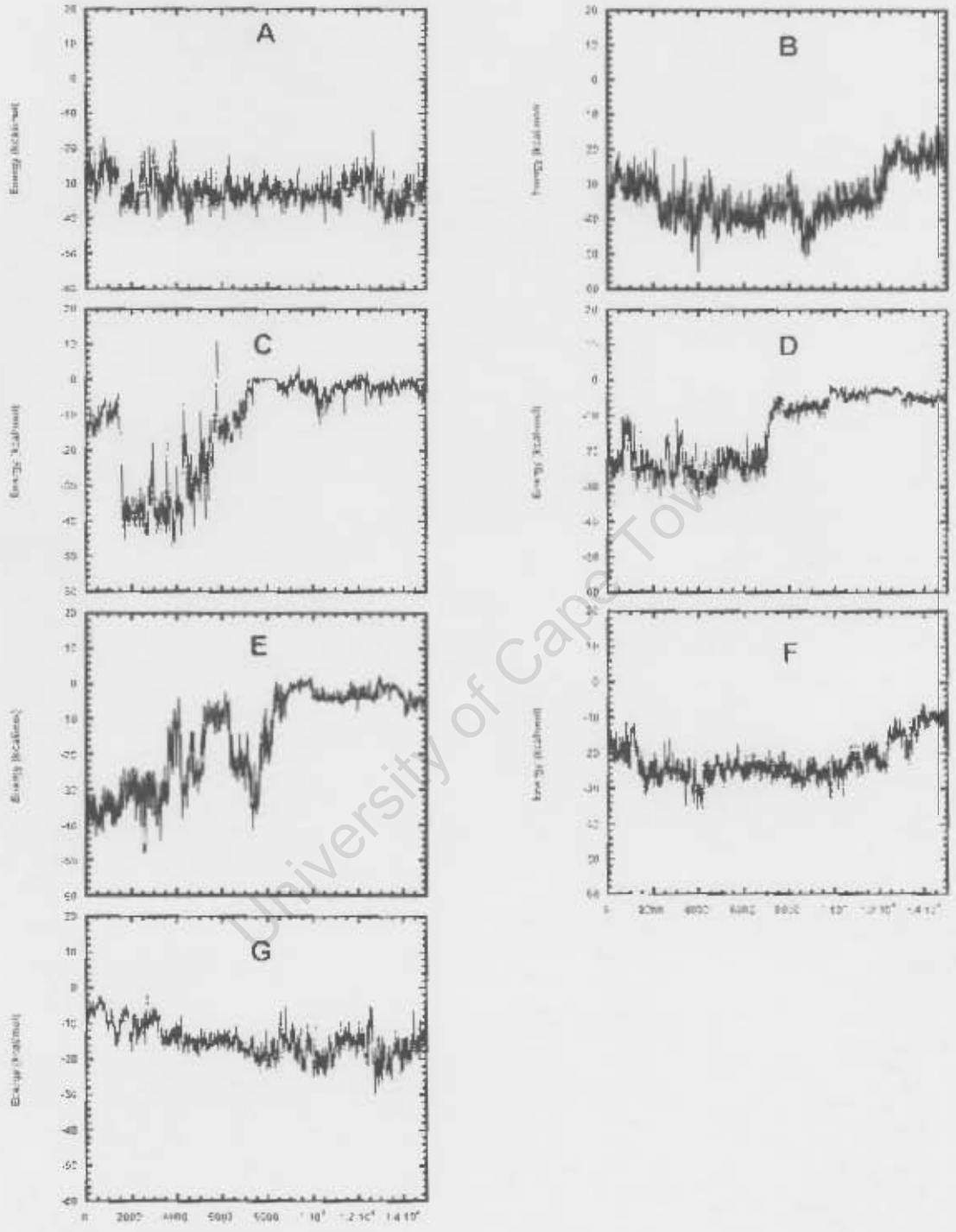


Figure 4.12 Region A - Interaction energy time series.

4.4.2.2 Binding of Guanine

Guanine is the second in the inhibition hierarchy according to both free energy results from this study and the experimental IC_{50} values. Binding of guanine to RTA is very similar to that of 9OG. H-bond analysis showed that there is only a single H-bond involved in binding of guanine, which is formed between N3 and amine hydrogen of VAL 81. Therefore, it is apparent that the primary mode of binding of guanine is vdW and /or electrostatic interactions. Guanine interacts with RTA only via region A. and this interaction behaviour is very similar to that of 9OG and to the rest of the inhibitors chosen for this study. In the binding of guanine, the residues TYR 80, VAL 81, ILE 172 and ARG 134 play significant roles.

The binding of guanine is predominantly by electrostatic interactions. The total interaction energy of guanine with region A is -33.32 kcal/mol and -21.82 kcal/mol of that is due to electrostatic interactions (Table 4.6). The time series of guanine – region A (Figure 4.12.B) shows that interaction of guanine with this region is steady. As seen with binding of 9OG, the involvement of ARG 134 is significant in binding of this inhibitor.

4.4.2.3 Binding of 9DG, 8M7DG and 7DG

Inhibitors 9DG, 8M7DG and 7DG are the weakest among the selected RTA inhibitors. Based on the free energy perturbation results they can be ranked as 9DG > 8M7DG > 7DG in inhibition power and this ranking is in perfect agreement with the experimental results. These inhibitors show ability to bind to RTA via interactions with region A. Furthermore, as observed with the inhibitors 9OG and guanine, none of these inhibitors interact with region B. The interaction energies of these inhibitors with region A are significantly less when compared to that of 9OG and guanine. Therefore, a clear relationship between their lower inhibition power and weak interactions with region A can be identified.

Moreover, it can be seen from interaction energy time series of these ligands (Figure 4.12 C, D,E for 9DG ,8M7DG and 7DG respectively) their interaction with region A is diminishing and approaching zero with time.

Therefore it can be conclusively stated that there exists a relationship between the interactions with region A and inhibition power of RTA inhibitors. It is attempted to rationalise this argument later in this discussion.

RMSD (for the ligand) analysis, distance analysis (distance from COG of TYR 80 to COG of the 6-membered ring of the ligand) and visual inspection of trajectories confirm that these inhibitors drifted away from the binding site with time. Therefore, caution was exercised when using and interpreting the average interaction energies presented in Table 4.6.

4.4.2.4 Binding of Adenosine (Substrate)

The binding orientation of adenosine was significantly different from that of the other ligands, possibly due to the presence of the ribose ring. The orientation of the base-ring of the ligands with respect to the regions A and B are shown in Figure 4.13. The special orientation of adenosine allows it to make more interactions with region B than any other inhibitor.

The binding of adenosine occurs primarily via non-bonded interactions. H-bond analysis showed that there is only a single H-bond formed between RTA and the substrate, which was found between N1 and amine hydrogen of VAL 81. The total average interaction energy of the substrate with ricin was -57.47 kcal/mol and -26.98 kcal/mol out of that total energy originates from the interactions with region B. The residues TYR 80, VAL 81 and ILE 172 are the key contact points in region A for adenosine binding. In region B, ARG 180 GLU 208 and TRP 211 were identified as the key residues. The vdW interactions between the substrate and the region B are negligible (-2.28 kcal/mol out of -26.98 kcal/mol) whereas the electrostatic interactions are quite significant (-24.70 kcal /mol out of -26.98 kcal /mol).

The interaction with GLU 208 is remarkable since it contributes to over 70% of the electrostatic interactions with region B. Therefore, it is clear that electrostatic interactions with region B are significant in binding of adenosine substrate to RTA. It is possible for adenosine to interact with region B simply via the sugar moiety when it is in a "parallel" orientation to the regions A and B (Figure 4.13(a)). The absence of the sugar ring in other inhibitors imposes a size constraint on them which prevents the interaction with region B.

Like other inhibitors, the substrate also interacts with region A. As mentioned before, the key points of interactions with region A are residues; TYR 80, VAL 81 and ILE 172. The interaction of substrate with this region is more vdW than electrostatic (vdW =-13.19 kcal/mol, elec.= -9.00 kcal/mol). This observation is different from the binding nature of the other inhibitors where electrostatic interactions are always greater than the vdW interactions. However, the time series of region A–substrate (Figure 4.12 F) interaction energy shows that the interaction with region A is continuous.

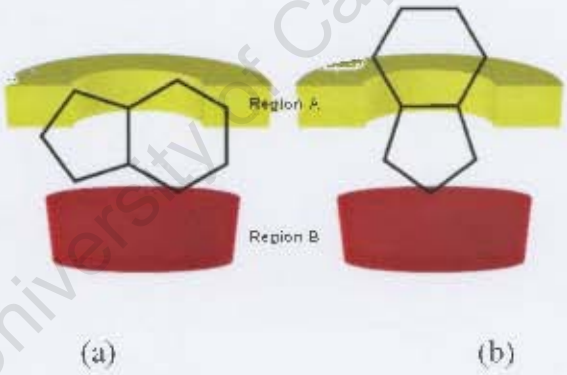


Figure 4.13 Schematic representation of the orientation of bicyclic moiety of the ligands in the RTA active site.

(a) - Orientation for substrate and product. (b) – Orientation for all other ligands.

4.4.2.5 Binding of Adenine base (Product)

Binding of adenine base is very similar to that of the substrate. As observed with the other ligands, there is no significant H-bonding involved in binding of this ligand. However, a single H-bond was observed between amine hydrogen of VAL 81 and N1 of the product. Therefore, it strongly suggests that binding of adenine base to RTA occurs primarily via vdW and electrostatic interactions.

The residues TYR 80 , VAL 81 are TYR 123 are the key residues involved in binding of product. The involvement of TYR 123 is noteworthy as it is only involved in binding of few ligands (9OG , guanine and 8M7DG). The interaction of the product with TYR 123 is mainly by Van der Waals interactions. Moreover, the majority of region A-product interaction is vdW, as in binding behaviour of the substrate. The time series of region A–product interaction (Figure 4.12G) shows that the product has continuous interaction with region A.

The interaction of product with region B is remarkable and this binding behaviour resembles that of the substrate. The vdW interactions with region B are negligible (-0.57 kcal/mol) whereas the electrostatic interactions are significant (-9.43 kcal/mol). The main interaction point for product to region B is residue ARG 180. This residue is proposed to be the acid (leaving group activator) in the proposed mechanism. The observed involvement of ARG 180 in binding of substrate (previous section) and product is corroborative of the anticipated role of ARG 180 in the proposed mechanism.

4.5 Rationalising the Binding Behaviour

It was attempted to understand the variation in the binding behaviour of different ligands (inhibitors, substrate and the product) using a selected set of analytical techniques. They are; analysis of the RMSD of the binding pocket, analysis of the movements of ligands about the centre of mass of the binding pocket and analysis of the electrostatic compatibility between the binding pocket and the ligands.

Time series of the root mean square deviation (RMSD) of the binding site was calculated over the binding of different ligands. These results are presented in Figure 4.14. According to the analysis of RMSD there are no significant structural changes occurring in the binding site due to the binding of different ligands. The changes observed in 9OG and adenine base were not taken into account seriously as they did not follow a periodic pattern. Therefore, the binding pocket was considered as relatively "rigid".

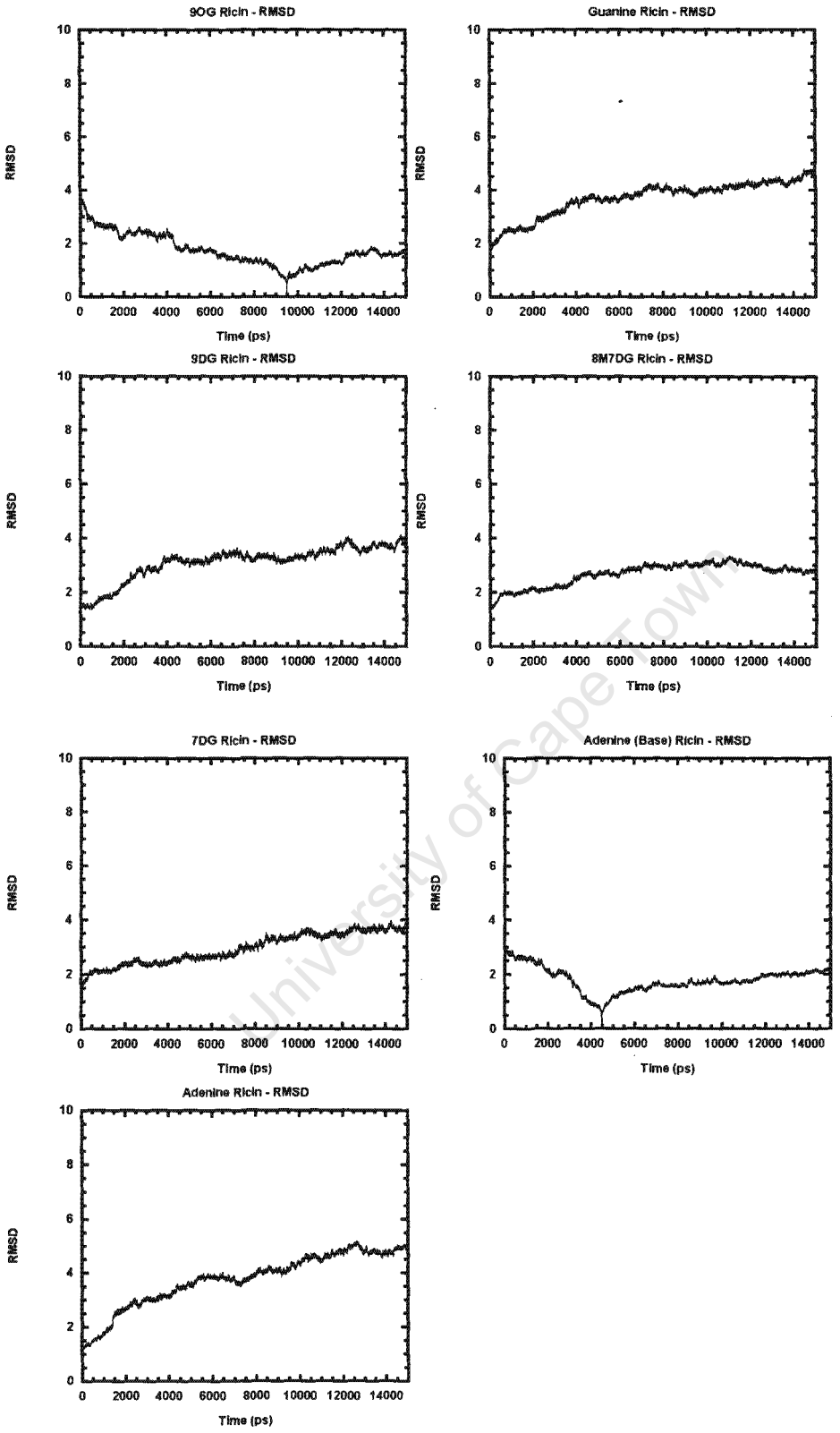


Figure 4.14 RMSD of the RTA binding site over the binding of different ligands.

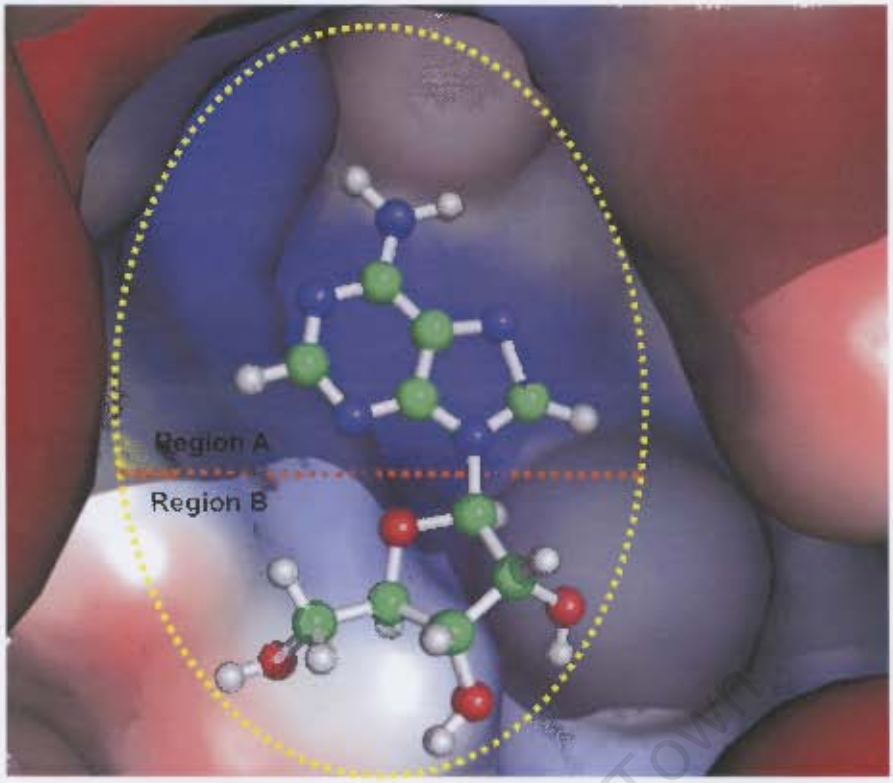


Figure 4.15 Electrostatic potential surface of RTA binding pocket

The electrostatic potential surface (EPS) of the binding pocket of RTA is shown in Figure 4.15 and regions A and B are demarcated with dashed lines. The highly positive potentials are represented in blue whereas red represents the highly negative areas. The EPS of the RTA binding pocket clearly shows that the region A is remarkably positive in its electrostatic potential.

The electrostatic potential contour maps of ligands were generated in order to investigate the electrostatic complementarity between the binding pocket and the ligands. An in-house program was used to generate the three dimensional electrostatic potential grid and the contour maps were generated using the Slicer program. The electrostatic contour maps of the selected inhibitors and the product (adenine base) are presented in Figure 4.16.

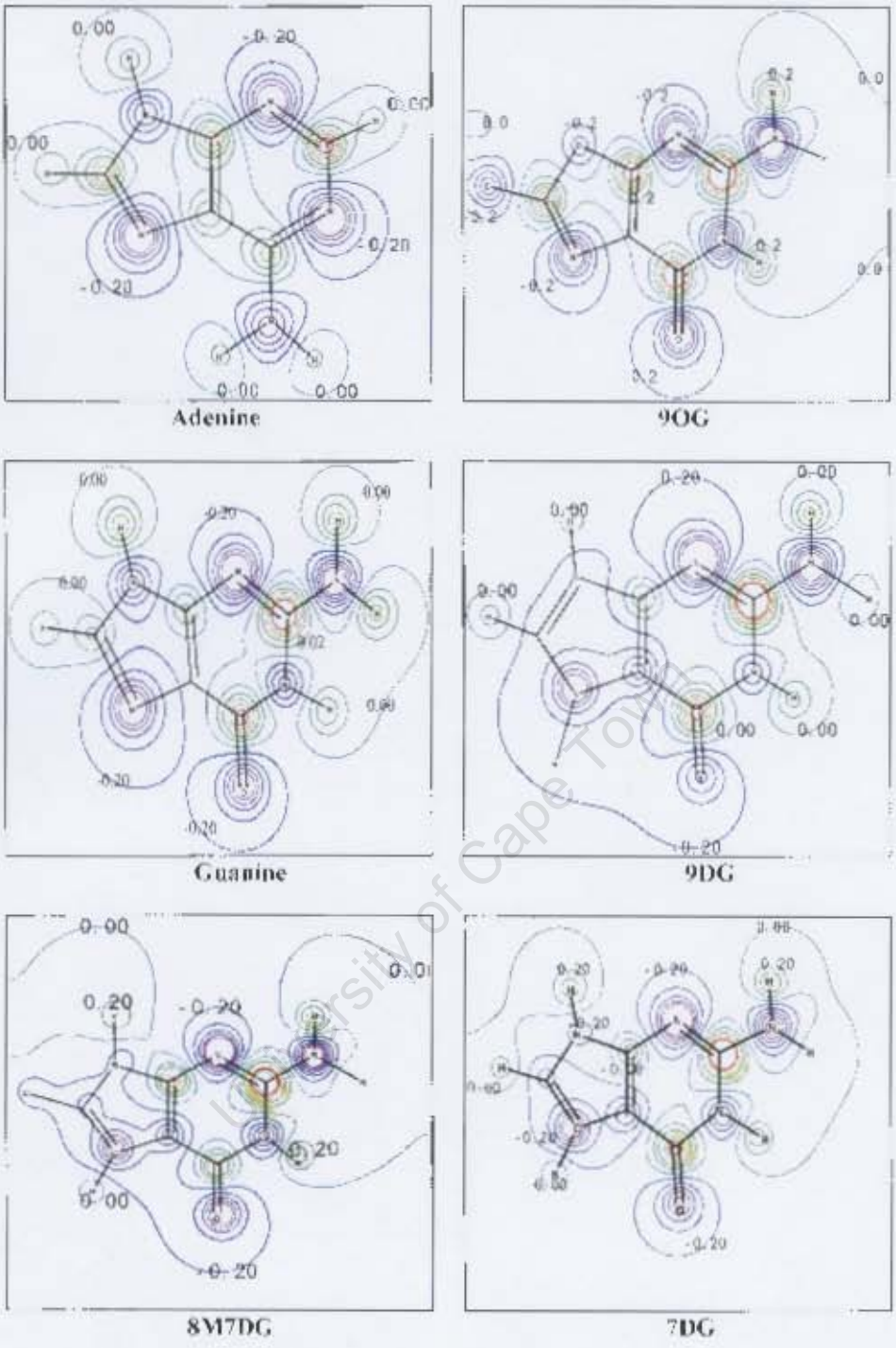


Figure 4.16 Electrostatic potential contour maps of ligands.

According to the EPS of the binding site, the ligands with sufficient number of negative "hot-spots" on their EPS must interact with region A in a quite significant manner.

The electrostatic contour map (ECM) of 9OG shows that there are five localised minima at the positions O6, N7, C81, O9 and N3 (key to table 4.3 is used as the numbering key). Moreover, these minima (-0.20) in the electrostatic potential (negative hot-spots) are easily accessible. Therefore, it can be concluded that the high electrostatic compatibility between the inhibitor 9OG and the RTA binding pocket's region A makes 9OG a powerful RTA inhibitor. The distance from the centre of mass of the binding pocket (region A + region B) and the centre of mass of 9OG remains constant throughout the simulation length of 15ns (Figure 4.17). Therefore previously explained steady interaction of 9OG with region A can be comprehended.

Three localised negative areas around the atoms O6, N7 and N3 can be identified on guanine based on its ECM . These sites show the easy access to them and do not have any shielding effect from the adjacent positive sites. The lesser number of negative sites (3) on guanine when compared to the number of negative sites (6) on 9OG suggest that guanine must have less electrostatic interactions with RTA. However, according to the interaction energy values shown in Table 4.6, the electrostatic interaction of guanine with region A is -21.82 kcal/mol and it is 4.27 kcal/mol larger than that of 9OG. The wider/inflated negative sites which are easily accessible on guanine can be accounted for this observation despite the lesser number of sites. Owing to the powerful electrostatic interactions, this inhibitor remains inside the binding pocket throughout and therefore, acts as a powerful inhibitor. Guanine's unperturbed stay in the binding site is reflected in the centre of mass (active site) – to-centre of mass (ligand) distance analysis shown in Figure 4.17.

9DG has two distinct negative sites, one localised around N3 and the other a diffused site in the region of the atoms O6, C5, N7 , C8 and C9. The negative site that is localised around N3 is easily accessible and therefore must play a significant role in binding to RTA. However, the diffused site is not easily accessible from certain points. The positive fields on the hydrogens at C9, C8 and N7 shield the underlying diffused negative region and therefore prevent access by any "negative seeking" residues in the binding pocket and this accounts for the weaker binding of 9DG in contrast to 9OG and guanine. The distance analysis (Figure 4.17) shows that this inhibitor is mobile / less settled inside the binding pocket and it has drifted away from the centre of the binding pocket by about 2.00 Å within the 15ns duration simulation time.

The distribution of the negative sites on 8M7DG is very similar to that of 9DG. However, both its localised negative site on N3 and the diffused site around O6, C5, C7, C8 and N9 are emaciated in contrast to these sites on 9DG. Specially, the shielding of the diffused site by the hydrogen atoms on C7 and N9 are much larger than the shielding seen in 9DG. Also, the localised negative site on N3 is shielded by the hydrogens on N7 and N2. Therefore, 8M7DG shows even less electrostatic compatibility with region A in comparison to 9DG, and this accounts for its lower inhibition power. As a result of the poor interaction with the binding site, this inhibitor moves out of the binding site gradually and this movement is reflected in the distance analysis results (Figure 4.17).

Inhibitor 7DG has three electronegative regions (Figure 4.16). Two of them are localised on N3 and O6 and the other is diffused around C5, C7 and C8. However, except for the site on O6, the others are highly shielded by the positive fields of the nearby hydrogens. As a result 7DG has only a single easily accessible negative site and this makes it the weakest inhibitor. As a result of extremely low electrostatic compatibility with the binding site, 7DG does not stay in the binding site. This can be seen from the distance analysis results (Figure 4.17).

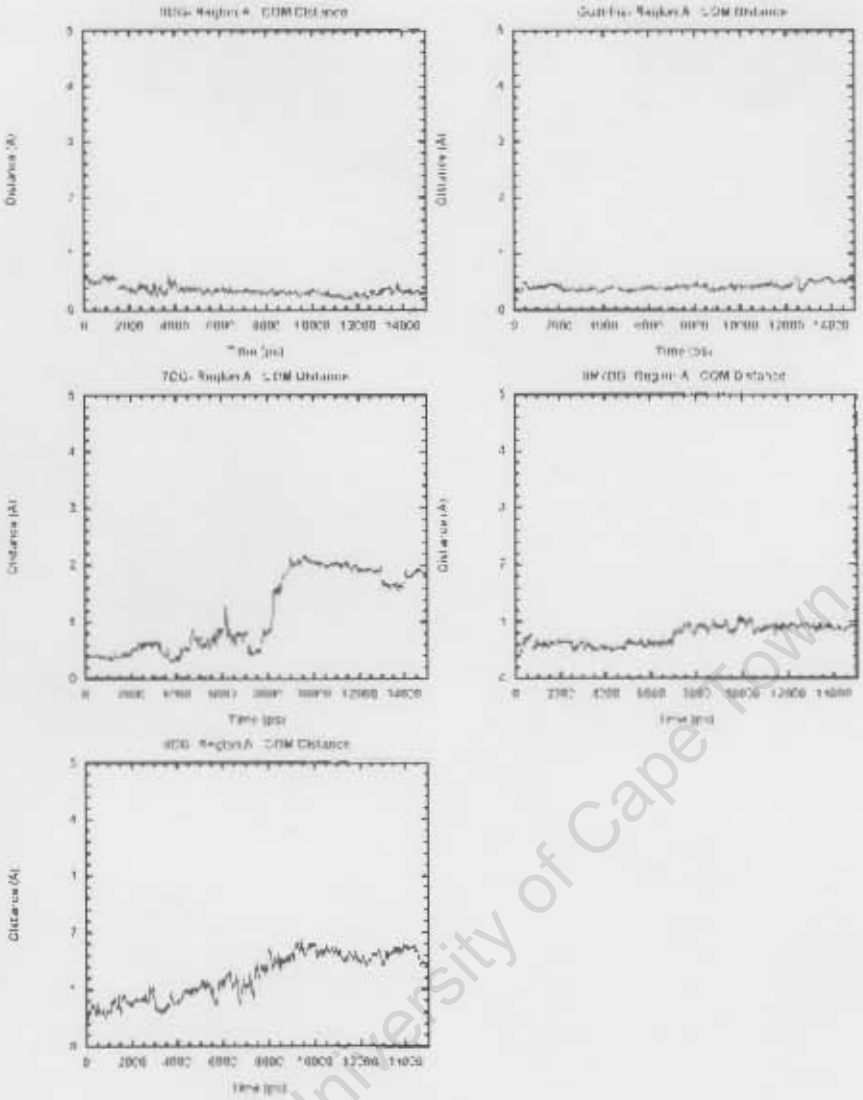


Figure 4.17 Distance between centre of mass of the binding pocket to centre of mass of the ligand.

4.6 Conclusions

This study has provided more insight into the understanding of the RTA binding pocket and its behaviour towards the binding of guanine-like inhibitors. In this study two distinct areas of RTA binding pocket were identified and the key amino acids / enzymatic contacts were clearly identified. The mode / nature of binding of selected inhibitors, reactants and products were clearly identified. Moreover it was found that the previously proposed idea of pi stacking of ligands between TYR 80 and TYR 123 is very highly unlikely to occur. Hydrogen bonding was not found to play a significant role in the binding of the selected inhibitors.

The results of this study indicate that RTA –ligand binding occurs mainly via van der Waals and electrostatic interactions. Therefore, electrostatic compatibility with the binding site (especially region A) is an essential feature of a good inhibitor. It was also seen from the results of this study, interaction with region B is significant in the binding process (as with substrate) as it contributes to very high electrostatic interactions.

Therefore, it can be concluded that the guanine-like inhibitors can still act as RTA inhibitors when there is a sufficient degree of electrostatic compatibility with region A, while having the correct shape. It is proposed that any better RTA inhibitor must posses; (1) a higher degree of electrostatic compatibility with region A (2) an ability to interact with the region B effectively. The guanine-like inhibitors are therefore not suitable candidates for further pharmacophore developments as they do not satisfy the 2nd condition mentioned above.

CHAPTER 5

Conclusion and Future Work

The primary objective of this thesis was to rationalise the inhibitory effect of the guanine-based ricin inhibitors. In order to achieve this objective, free energy perturbation methods were employed to rank the inhibitors. This ranking showed exceptional agreement when compared to the experimental inhibition results obtained from IC_{50} values.

Secondly, MD simulations were used to identify key amino acids present in the RTA binding site. This identification was then used improve understanding of the mode of binding of the different ligands to it. Further analysis of the MD runs revealed that electrostatic and van der Waals interactions play the dominant role in the binding of the substrate and selected inhibitors to RTA. These interactions also lead to the identification of two distinct regions of the RTA binding site (Region A and B).

Since electrostatic interactions play a significant role in binding, the electrostatic compatibility between the binding site and the ligands was intensively examined. These results were then used to interpret the variation of the binding behaviour for different ligands. Accordingly, the inhibitors with an overall negative electrostatic potential were found to be more powerful owing to the high electrostatic compatibility with the region A of RTA. A direct relationship between the region A compatibility and inhibitory activity was established.

The importance of the region B in RTA-ligand binding was discovered through the binding of the substrate and the product. The binding of the substrate to RTA is extensively supported by electrostatic interactions with the region B. Therefore, the region B was identified as the substrate-specific region of the RTA binding site. However, all selected inhibitors failed to interact with the region B and were thus poor inhibitors.

In conclusion, it is suggested that the guanine-like RTA inhibitors (i.e. the inhibitors that resemble the product adenine) can bind relatively strongly to RTA only if they posses a higher degree of electrostatic compatibility with the region A of the binding site. However, since guanine-like inhibitors lack interaction with the region B it is proposed that they are not good candidates for further pharmacophore development.

Development of better inhibitors for RTA requires a complete understanding of its reaction mechanism. The reaction pathway (mechanism) is the path that describes the movement of a system (reactants) from one energy minimum to another (products) via a saddle point. It is a generally accepted fact that a chemical reaction is simulated ideally with quantum mechanical methods. However, for large systems such as proteins when all atoms are explicitly represented, quantum mechanical calculations become practically impossible. As a result, numerous alternative methods been developed to model chemical reactions. It is possible to identify three major approaches in these methods: They are the pure empirical approach, quantum mechanical – molecular mechanical (QM/MM) hybrid model and the Car-Parrinello approach [38]. In future studies the reaction mechanism of RTA is to be explored by the most appropriate of these methods. Complete understanding of the reaction mechanism and the transition state structure is necessary to allow proposals of new pharmacophores for RTA.

References

1. Voet, D., Voet, J.G.W, P., in *Fundamentals of Biochemistry*. 1999, John Wiley & Sons,Inc. p. 663.
2. Chaplin, M.Bucke, C., *Enzyme Technology*. 1st ed. 1990: Cambridge Unversity Press.
3. Pelzer, H.Wigner, E., *Physical Chemistry B*, 1932. **15**: p. 445.
4. Fersht, A., *Strucuture and Mechanism in Protein Science*. 3rd ed. 1999: W.H. Freeman and Company; New York.
5. Korolev, V.G., *Russian Journal of Genetics*, 2005. **41**(6): p. 583.
6. Stirpe, F., *ToxiconHighlights in plant toxins*, 2004. **44**(4): p. 371.
7. Yeung, H.W., *Journal of Peptide Research*, 1988. **31**: p. 265.
8. Liu, R.S., *European Journal of Biochemistry*, 2002. **269**: p. 4746–4752.
9. Endo, Y., *Journal of Biological Chemistry*, 1987. **262**(12): p. 5908.
10. Miller, D.J., Ravikumar, K., Shen, H., Suh, J.K., Kerwin, S.M.Robertus, J.D., *Journal of Medicinal Chemistry*, 2002. **45**(1): p. 90.
11. Stillmark, H., *Uber Ricin, ein giftiges. Ferment aus den Samen von Ricinus communis L. und anderen Euphorbiaceen Dorpat Estonia*. 1888.
12. Yan, X., Hollis, T., Svinth, M., Day, P., Monzingo, A.F., Milne, G.W.A.Robertus, J.D., *J.Mol.Biol*, 1997. **266**: p. 1043.
13. Lin, J.-Y., *Cancer Research*, 1971. **31**: p. 921–924.
14. Sandvig, K.van Deurs, B., *Febs Letters*, 2002. **529**(1): p. 49.
15. Lord, J.M., *Biochemical Society Transactions*, 2003. **31**: p. 1260.
16. Endo, Y., *Journal of Molecular Biology*, 1991. **221**(1): p. 193.
17. Endo, Y., *Journal of Biological Chemistry*, 1987. **262**(17): p. 8128.
18. Yang, X., *Natural Structural Biology*, 2001. **8**: p. 968.
19. Chen, X.Y., Berti, P.J.Schramm, V.L., *Journal of the American Chemical Society*, 2000. **122**(8): p. 1609.
20. Chen, X.Y., *Journal of the American Chemical Society*, 2000. **122**(28): p. 6527.
21. Roday, S., Saen-Oon, S.Schramm, V.L., *Biochemistry*, 2007. **46**(21): p. 6169.
22. Yan, X., Hollis, T., Svinth, M., Day, P., Monzingo, A.F., Milne, G.W.A.Robertus, J.D., *Journal of Molecular Biology*, 1997. **266**(5): p. 1043.
23. Montfort, W., *Journal of Biological Chemistry*, 1987. **262**(11): p. 5398.
24. Rutenber, E., *Proteins*, 1991. **10**: p. 240.
25. Monzingo, A.F.Roberts, D.J., *Journal of Molecular Biology*, 1992. **227**: p. 1136.
26. Kim, Y., *Biochemistry*, 1992. **31**: p. 3294.
27. Haran, G., *Journal of Physics-Condensed Matter*, 2003. **15**(32): p. R1291.
28. Baryshnikova, E.N., *Protein Science*, 2005. **14**(10): p. 2658.
29. WANG, T., *Biophysical Journal*, 2005. **89**: p. 1.
30. DM, V., *BIOCHEMISTRY*, 2004. **43**: p. 3582.
31. DYSON , H.J., *Nuclear Magnetic Resonance of Biological Macromolecules Part C*, 2005. **394**: p. 299.
32. Chen, X.Y., Link, T.M.Schramm, V.L., *Biochemistry*, 1998. **37**(33): p. 11605.
33. Karplus, M., *Nature Structural Biology*, 2002. **9**(9): p. 646.
34. Adcock, S.A.McCammon, J.A., *Chem. Rev.*, 2006. **106**(5): p. 1589.
35. Berendsen, H.J., *Current Opinion in Structural Biology*, 2000. **10**: p. 165.

36. Gunsteren, W.F.Mark, A.E., *European Journal of Biochemistry*, 1992. **204**: p. 947.
37. Lewars, E., *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics.* 1st ed. 2004: Kluwer Academic Publishers.
38. Leach , A.R., *Molecular Modelling: Principles and Applycations*. 3rd ed. 2001: Prentice Hall; New Yourk.
39. Foresman, J.B., *Exploring Chemistry with Electronic Structure*. 2nd ed. 1996: Gaussian, Inc.
40. McQuarrie, *Quantum Chemistry*. 1s ed. 1983: University Science Books.
41. Brooks, B.R., Bruccloleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S.Karplus, M., *Journal of Computational Chemistry*, 1983. **4**(2): p. 187.
42. Allen, M.P., *Computer Simulation of Liquilds*. 1987: Oxford University Press, Oxford.
43. Haile, J.M., *Molecular Dynamics Simulation: Elemetary methods*. 1992: John Wiely & Sons Inc, New York.
44. McCammon, J.A., *Nature*, 1977. **267**: p. 585.
45. Field, M.J., *Journal of Computational Chemistry*, 1989. **11**(6): p. 700.
46. Verlet, L., *Physical Review*, 1967. **159**: p. 98.
47. Hockney , R.W., *Methods in Computational Physics*, 1970. **9**: p. 136.
48. Swope, W.C., *Journal of Chemical Physics*, 1982. **76**: p. 637.
49. Beeman, D., *Journal of Computational Physics*, 1976. **20**: p. 130.
50. Cramer, C.J.Truhlar, D.G., *Chemical Reviews*, 1999. **99**: p. 2161.
51. Simonson, T., *Current Opinion in Structural Biology*, 2001. **11**: p. 243.
52. Feig, M., *Current Opinion in Structural Biology*, 2004. **14**: p. 217.
53. Lazardis, C.Karplus, M., *Biophysical Chemistry*, 2003. **100**: p. 367.
54. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W.Klein, M.L., 1983. - **79**(- 2).
55. Mahoney, M.W.Jorgensen, W.L., *Journal of Chemical Physics*, 2000. **112**: p. 2000.
56. Berendsen, H.J.Grigera, J.R., *Journal of Physical Chemistry*, 1987. **91**: p. 6269.
57. Vriend, G., *Journal of Molecular Graphics*, 1990. **8**(1): p. 52.
58. Singh, U.C.Kollman, P.A., *Journal of Computational Chemistry*, 1984. **5**: p. 129.
59. Lee, C., Yang, W.Y.Parr, R.G., *Physical Review B*, 1988. **73**: p. 785.
60. Pearlman, D.A.Rao, B.G., *Freee Energy Calculations: Methods and Applycations*. 1st ed. Encyclopedia of Computational Chemistry, ed. P. van Rague` Schleyer. 1998: Jhon Wily & Sons:Chichester.
61. Becveridge, D.L., *Annual Review of Biophysics and Biophysical Chemistry*, 1989. **18**: p. 431.
62. Leah , A.R., *Molecular Modelling: Principles and Applycations*. 3rd ed. 2001: Prentice Hall; New Yourk.
63. Straatsma, T.P., *Chemical Physics Letters*, 1992. **196**(3-4): p. 297.
64. Kollman, P., *Chemical Reviews*, 1993. **93**(7): p. 2395.
65. Pearlman, D.A., *Journal of Chemical Physics*, 1989. **91**(12): p. 7831.
66. Pearlman, D.A.Kollman, P., *Journal of Chemical Physics*, 1989. **91**: p. 7831.
67. Hermans, J., *Journal of Physical Chemistry*, 1991. **95**: p. 9029.
68. Mitchell, M.J.McCammon, J.A., *Journal of Computational Chemistry*, 1991. **12**: p. 271.

69. Zwanzig, R.W., *Journal of Chemical Physics*, 1954. **22**: p. 1420.
70. Pearlman, D.A.e.a., *Computational Physcs Communication*, 1995. **91**: p. 1.
71. Brooks, B.R.e.a., *Journal of Computational Chemistry*, 1983. **4**: p. 187.
72. Gao, J., Kuczera, K.Tidor, B., *Science*, 1989. **244**: p. 1069.
73. Prevost, M., *Proc.Natl.Aad.Sci*, 1991. **88**: p. 10880.
74. Mark, A.E., *Freee Energy Perturbation Calculations*. 1st ed. Encyclopedia of Computational Chemistry, ed. P. van Rague` Schleyer. 1998: Jhon Wily & Sons:Chichester.
75. Straatsma, T.P., *Freee Energy Simulations*. 1st ed. Encyclopedia of Computational Chemistry, ed. P. van Rague` Schleyer. 1998: Jhon Wily & Sons:Chichester.
76. Cieplak, P., *Journal of the American Chemical Society*, 1988. **110**(12): p. 3734.
77. Jorgensen, W.L., *Journal of Chemical Physics*, 1988. **89**(6): p. 3742.
78. Chakravartula, S.V.S.Guttarla, N., *Natural Product Research*, 2008. **22**(3): p. 258.
79. Day, P.J., *Biochemistry*, 1996. **35**(34): p. 11098.
80. Chaddock, J.A.Roberts, L.M., *Protein Engineering Design and Selection Protein Eng.*, 1993. **6**(4): p. 425.
81. Mark A. Olson, *Proteins: Structure, Function, and Genetics*, 1997. **27**(1): p. 80.
82. Olson, M.A.Cuff, L., *Biophysical Journal*, 1999. **76**(1): p. 28.
83. Spackova, N.a.Sponer, J., *Nucleic Acids Research.*, 2006. **34**(2): p. 697.
84. Casey , R.M., *Bioinformatics in Structure-Based Drug Design*. 2006, Powell Media LLC.
85. Foloppe, N.MacKerell, A.D., *Journal of Computational Chemistry*, 2000. **21**(2): p. 86.
86. MacKerell, A.D.Banavali, N.K., *Journal of Computational Chemistry*, 2000. **21**(2): p. 105.
87. van Gunsteren, W.F.Berendsen, H.J.C., *Molecular Physics*, 1977. **34**: p. 1311.
88. Singh, U.C.Kollman, P.A., *Journal of Computational Chemistry*, 1984. **5**: p. 129.
89. McGaughey , G.B., *Journal of Biological Chemistry*, 1998. **273**(25): p. 15458.
90. Hunter, C.A., *Journal of American Chemical Society*, 1990. **112**: p. 5525.