University of Alberta

Modelling and Simulation of Carbohydrate Systems: from Solution to Protein Binding in the Gas Phase

by

Mikyung Seo ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Doctor of Philosophy**.

Department of Chemistry

Edmonton, Alberta
Fall 2008

Your file Votre référence
ISBN: 978-0-494-46420-5
Our file Notre référence
ISBN: 978-0-494-46420-5

Abstract

This thesis presents studies aimed at providing a deeper understanding of protein-carbohydrate recognition and binding process by means of theoretical modelling and computational simulations. This study focuses on the theoretical modelling of furanosides in particular. Furanoside ring systems can exist in a variety of conformers that are separated by low energy barriers. This conformational flexibility renders the modelling of such systems quite challenging. An earlier model had been shown to be reliable for simulation studies of more rigid pyranoside ring systems. Our new model, however, incorporates the effects of ring flexibility for better description of furanosides. It is applied to a monosaccharide, the methyl α-D-arabinofuranoside. The simulations results are compared to the results from NMR experiments in order to assess the validity of the model. For the study of the interactions between proteins and carbohydrates, an antibody single-chain fragment (scFv) and its native ligand, αGal[αAbe]αMan are investigated as a model system. The intermolecular hydrogen bonds within a *desolvated* protein-ligand complex are characterized. This work is inspired by recent mass spectrometry experiments where some specific interactions are found to be preserved from solution to gas phase. We provide a more complete picture of the nature of the binding interactions between the protein and the ligand through hydrogen bond analysis. This analysis method is also used to probe the interactions between crystallographic" water molecules and the complex. Through calculations of free energy profiles of the native complex and its mutants, the contributions of var-

ious specific protein-ligand interactions to the stability of the mutant complexes are observed. This study also allows us to quantify the strength of specific interactions.

Acknowledgements

The biggest thanks go to my supervisor, Prof. Pierre-Nicholas Roy for his generous support, patience, and encouragement throughout the entire process of my graduate career. He has been a wonderful advisor, always available for discussion, helping my research through many insightful discussions, showing great care for his students professional development, and considering their different personal situations. Specially, I am deeply grateful to him for being considerate when I had two babies, for allowing me flexible work hours due to busy life with a family, and for providing additional funding during my maternity leaves. I couldn't have hoped for a better supervisor.

I would especially like to acknowledge Dr. Nicholas Blinov for his help and many valuable discussions. In particular, he showed exceptional patience when helping me with the Bioinformatics course. I also thank Prof. John Klassen and Dr. Elena Kitova who have collaborated with me on research related to the protein-carbohydrate complexes. I would like to thank Dr. Bilkiss Issack, Yoonjung Huh, Dr. Norberto Castillo, Dr. Javier Cuervo and Hashem Taha for their comments and advices on this thesis. Special thanks go to Dr. Bilkiss Issack for helping me with my course works and research projects, especially in the last stages of thesis. I am also grateful to my Ph. D. committee members for important comments and suggestions to that improved this thesis.

I thank the following former and present Roy group members (in the alphabetical order): Dr. Norberto Castillo, Dr. Javier Cuervo, Dr. Robert Ganzynkowicz, Kyle Greene, Yoonjung (Yoonie) Huh, Dr. Bilkiss Issack, Dr. Yong Dong (Lucy) Liu, Yuan Ma, Paul Moffat, Yalina Tritzant and Stephanie Wong for their support and advice in research in general, and for their wonderful friendship. They have made my time at school enjoyable, comfortable and less stressful.

Most of all I would like to thank my family, from parents to siblings, who have helped me throughout this process with their love, never-ending support, and prayers. Special thanks go to my beautiful children, Joochan and Yechan, the non-academic achievements of my graduate school years. They have given me the joy and strength to overcome obstacles in difficult times, and have showed understanding and patience towards a busy mother in my final year Finally, I thank Jonghwa for his support, love, patience, sacrifice and prayers, day and night, throughout our marriage.

Contents

List of Figures

List of Tables

List of Abbreviations

NMR	*Nuclear Magnetic Resonance*
ScFv	*Single-Chain Fragment*
BO	*Born-Oppenheimer*
PES	*Potential Energy Surface*
MD	*Molecular Dynamics*
MM	*Molecular Mechanics*
DNA	*Deoxyribonucleic Acid*
RNA	*Ribonucleic Acid*
MC	*Monte Carlo*
PMF	*Potential of Mean Force*
WHAM	*Weighted Histogram Analysis Method*
AMBER	*Assisted Model Building and Energy Refinement*
CHARMM	*Chemistry at HARvard Macromolecular Mechanics*
OPLS	*Optimised Potential for Liquid Simulations*
GROMOS	*GROningen MOlecular Simulation*
HSEA	*Hard Sphere Exo-Anomeric*
GLYCAM	*Oligosaccharide/glycoprotein force field*
AG	*ArabinoGalactan*
LAM	*LipoArabinoMannan*
HF	*Hartree-Fock*
TIP3P	*Transferable Intermolecular Potential 3 Point*
ESP	*ElectroStatic Potential*

RESP	*Restrained ElectroStatic Potential*
RMSD	*Root Mean Square Displacement*
DFT	*Density Functional Theory*
ES	*Electrospray Ionization*
MS	*Mass Spectrometry*
FT-ICR	*Fourier-Transform Ion Cyclotron Resonance*
BIRD	*Blackbody Infrared Radiative Dissociation*
FGR	*Functional Group Replacement*
HBLTCF	*Hydrogen Bond Lifetime Correlation Function*
FEP	*Free Energy Perturbation*
TI	*Thermodynamic Integration*
TST	*Transition State Theory*
VTST	*Variational Transition State Theory*

Chapter 1

Introduction

1.1 Context

"*The secret of life is molecular recognition; the ability of one molecule to 'recognize' another through weak bonding interactions.*" Linus Pauling is reported to have said these words at the 25th anniversary of the Institute of Molecular Biology at the University of Oregon [1]. Molecular recognition by specific targets lies at the heart of life processes. It refers to the non-covalent specific interactions between two or more biological molecules: protein-carbohydrate, receptor-ligand, antigen-antibody, DNA-protein and many other interactions. It is well established that carbohydrates play a role in important biological recognition processes. The molecular interactions involved in the recognition of carbohydrates by proteins mediate a broad range of biological activities, such as cell growth, the immune response and bacterial infection. The elucidation of the mechanisms that govern how oligosaccharides are accommodated in the binding sites of lectins, antibodies, and enzymes is currently a topic of major interest [2, 3, 4]. Therefore, characterizing the structure and energetics, and understanding the interactions of protein-carbohydrate complexes are key to understanding many biological functions and successful drug design.

A variety of experimental techniques including X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy, microcalorimetry, and site-directed mutagenesis have provided a great deal of information on the structural and energetic principles underlying these important protein-carbohydrate interactions [5].

Detailed information on the structure of the complexes is however difficult to obtain since the usually high molecular weight of proteins has prevented their direct studies by means of NMR spectroscopy [6]. In addition, the potential conformational flexibility exhibited by some oligosaccharides impede the growth of crystals, both in the free and complexed form [7]. Unaffected by such experimental restrictions, computational approaches offer a promising alternative means of exploring the conformational preferences of flexible carbohydrates, and of understanding protein-carbohydrate recognition processes at a microscopic level. During the past decade, developments in computational methodologies have been made in the area of oligosaccharide conformational analysis and in the field of protein-oligosaccharide interactions. Continually evolving computational methods combined with increasing computer power not only broaden the range of feasible applications, but also open up detailed observations of protein-ligand interactions.

The computational investigation of molecular systems requires a theoretical description of the molecules of interest. The accuracy of the description or model, most often defines the reliability of the computed results. The conformational diversity and complexity of carbohydrates makes them a challenging class of molecules to model. Furanosides, in particular, are highly flexible five-membered ring molecules whose biological functions are affected by their conformations [8, 9, 10, 11]. In the first part of the thesis, our developmental efforts in carbohydrate modelling are presented. The notable difference that distinguishes this new approach from an existing model for furanosides [12, 11] is the consideration of the inherent flexibility of furanoside rings. The performance of the model is then illustrated through the elucidation of the solution conformation of carbohydrates containing furanose rings.

The second area of study presented in this thesis aims at investigating the factors affecting the intrinsic binding interactions between protein and carbohydrates at the molecular level. Computer simulations are carried out for a gaseous protein-trisaccharide complex consisting of a genetically engineered single chain variable fragment, scFv, of the monoclonal antibody Se155-4 and its native trisaccharide ligand, αGal[αAbe]αMan, in which Gal, Man, Abe stand for galactose, mannose, abequ-

ose (3,6-dideoxy-D-hexose), respectively [13]. The trisaccharide ligand, which is an epitope of the *Salmonella* group B O-antigen, represents an important disease markers and a target for therapeutic antibodies [14]. Due to the limited availability of simulation data for the antibody-carbohydrate complex, a full understanding of the recognition (or binding) process has not yet been achieved to date from a theoretical point of view. One of the aims of our work is to contribute to this understanding by offering a molecular view of the process.

This introduction is organized as follows. We start in Section 1.2 with a description of the Born-Oppenheimer approximation, one of the most important approximations employed when solving Schrödinger's equation for complex systems containing more than one or two electrons. Section 1.3 offers an introduction to some useful concepts of statistical mechanics. The following section gives an overview of classical Molecular Dynamics (MD) and the typical properties that can be calculated (Section 1.4). Lastly, challenges in carbohydrate modelling and the investigation of protein-carbohydrate interactions are discussed in Section 1.5. The development of computational methodology to model protein-carbohydrate interactions is described therein.

1.2 Born-Oppenheimer Approximation

At the beginning of the twentieth century, experimental evidence suggested that atomic particles were also wave-like in nature. It is reasonable to assume that a wave equation could explain the behaviour of atomic particles. The Schrödinger equation is the wave equation used in quantum mechanics to describe the behaviour of particles.

The time-independent Schrödinger equation is given by the following eigenvalue problem,

$$\hat{H}\Psi(\mathbf{r},\mathbf{R}) = E_{tot}\Psi(\mathbf{r},\mathbf{R}) \ , \tag{1.1}$$

where $\mathbf{R}$ and $\mathbf{r}$ are position vectors of the nuclei and electrons, respectively. $\Psi(\mathbf{r},\mathbf{R})$ is the wave function, $\hat{H}$ is the Hamiltonian and E_{tot} is total energy.

For a molecular system, the Hamiltonian is given as a sum of five terms:

$$\hat{H} = -\sum_{A} \frac{\hbar^2}{2m_A}\nabla_A^2 - \frac{\hbar^2}{2m_e}\sum_{i}\nabla_i^2 - \frac{e}{4\pi\varepsilon_0}\sum_{i,A}\frac{Z_A}{|r_i - R_A|} + \frac{e^2}{4\pi\varepsilon_0}\sum_{i<j}\frac{1}{|r_j - r_i|} + \frac{1}{4\pi\varepsilon_0}\sum_{A<B}\frac{Z_A Z_B}{|R_B - R_A|} , \tag{1.2}$$

where A and B refer to nuclei and i and j refer to electrons, r denotes electronic positions, and R the nuclear positions. The molecular system consists of electrons with the electronic mass m_e and nuclei with mass m_A. Electrons and nuclei have charges e and Z, respectively. In this Hamiltonian, the first two terms represent the kinetic energy and the last three terms potential energy. The first term is associated with the kinetic energy of the nuclei, and the second term with the kinetic energy of the electrons. The third term is the attractive Coulombic interaction between nuclei and electrons. The last two terms represent the energy associated with the repulsion resulting from like-charge interactions, namely the electron-electron and the nuclear-nuclear interactions [15].

Eq. (1.2) can be written more compactly as

$$\hat{H} = \hat{T}_N(\mathbf{R}) + \hat{T}_e(\mathbf{r}) + \hat{V}_{eN}(\mathbf{r},\mathbf{R}) + \hat{V}_{ee}(\mathbf{r}) + \hat{V}_{NN}(\mathbf{R}) , \tag{1.3}$$

where $\hat{T}$ are the kinetic energy and $\hat{V}$ is the potential energy operator, respectively. The particles involved in each operator are indicated by subscripts, e for electrons and N for nuclei.

Ideally, the Schrödinger equation should be solved for the wave function $\Psi(\mathbf{r},\mathbf{R})$, a function described by all position variables, $\mathbf{R}$ and $\mathbf{r}$. Solving the Schrödinger equation for systems of higher complexity than an atom with one electron is unfeasible to carry out in practice because many degrees of freedom need to be taken into account. Thus, various approximations need to be imposed for such molecular systems.

One of the most important and fundamental approximations used is called the Born-Oppenheimer (BO) approximation, developed by Max Born and Robert J. Oppenheimer [16] in 1927. The BO approximation separates electronic and nuclear motion based on the idea that the nuclear mass is so much larger than the mass of an

electron that the nuclei are basically "fixed" particles. Let us think classically about this difference in mass. If two particles attract in some way, and one is much heavier than the other, the light particle will simply follow the heavy particle wherever it goes, and it will respond instantaneously to changes in the position of the heavy particle. It is, therefore, safe to consider the nuclei as being fixed with respect to electronic motion, and so the true wave function $\Psi(\mathbf{r}, \mathbf{R})$ can be factorized into a product of two functions $\psi(\mathbf{r}; \mathbf{R})\phi(\mathbf{R})$, where $\phi(\mathbf{R})$ depends on the nuclear position variables, and $\psi(\mathbf{r}; \mathbf{R})$ the electronic ones (depending only parametrically on the nuclear variables). With this assumption in hand, the time-independent Schrödinger equation can be re-written in terms of two separate equations. The first involves the motion of the electrons for given nuclear positions:

$$\hat{H}_{el}\psi(\mathbf{r}; \mathbf{R}) = E_{el}(\mathbf{R})\psi(\mathbf{r}; \mathbf{R}) \ , \tag{1.4}$$

where

$$\hat{H}_{el} = \hat{T}_e(\mathbf{r}) + \hat{V}_{eN}(\mathbf{r}, \mathbf{R}) + \hat{V}_{ee}(\mathbf{r}) + \hat{V}_{NN}(\mathbf{R}) \ . \tag{1.5}$$

The electronic energy eigenvalue E_{el} depends on the given positions $\mathbf{R}$ of the nuclei. Varying these nuclear positions $\mathbf{R}$ in small steps and repeatedly solving the electronic Schrödinger equation Eq. (1.4), one obtains the electronic energy E_{el} as a function of $\mathbf{R}$. This electronic energy $E_{el}(\mathbf{R})$ is used as the potential energy surface (PES) in the Schrödinger equation, and it can be described the motion of the nuclei on the PES:

$$\hat{H}_{nuc}\phi(\mathbf{R}) = E_{tot}\phi(\mathbf{R}) \ , \tag{1.6}$$

where

$$\hat{H}_{nuc} = \hat{T}_N(\mathbf{R}) + E_{el}(\mathbf{R}) \ . \tag{1.7}$$

In Eq. (1.7), the Hamiltonian $\hat{H}_{nuc}$ for the nuclear motion equals the nuclear kinetic energy operator $\hat{T}_N(\mathbf{R})$ plus the electronic energy $E_{el}(\mathbf{R})$. In principle, Eq. (1.4) should be solved for $E_{el}(\mathbf{R})$, and then Eq. (1.7) for the nuclear motion. We "just" have to perform a series of electronic structure calculations for given nuclear positions using *ab initio* quantum chemistry methods. But finding the solutions to

Eqs. (1.4) and (1.7) requires a large amount of computation. Thus, an empirical method to fit $E_{el}(\mathbf{R})$ (also called a force field) is often employed. A classical force field consists of analytical functional forms describing the interactions between atoms in a system and parameters in these functional forms. The force fields are typically expressed as a sum of bonded interactions corresponding to chemical bonds, bond angles and dihedral angles, and non-bonded interactions associated with van der Waals forces and electrostatic charges. This will be discussed further in Section 1.5.1.

Since the nuclei are relatively heavier than electrons, the quantum mechanical effects of the nuclei are often insignificant, and Eq. (1.7) can be replaced with classical equations of motion (*i.e.*, the Newton's equation of motion, $\mathbf{F}_i = m_i\mathbf{a}_i$):

$$-\frac{dE_{el}(\mathbf{R})}{d\mathbf{R}_i} = m\frac{d^2\mathbf{R}_i}{dt^2} \ , \tag{1.8}$$

where $\mathbf{F}_i = -\frac{dE_{el}(\mathbf{R})}{d\mathbf{R}_i}$ is the force on nucleus i, and m_i and $\mathbf{a}_i$ are the mass and the acceleration of the nucleus, respectively [17]. Solving this equation for the classical motion of nuclei on a single PES $E_{el}(\mathbf{R})$ (*i.e.*, associated with a single electronic state) is called classical molecular dynamics (MD). Classical MD studies the time evolution of a system consisting of nuclei or atoms according to classical mechanics. If we are not interested in the time evolution of a system, $E_{el}(\mathbf{R})$ can be used to calculate static properties such as equilibrium structures, transition states and relative energies. This is called molecular mechanics (MM).

It is important to note that there are several assumptions in classical MD: First, by invoking the BO approximation, the electronic coordinates are averaged over the electronic wave function, so the nuclei move in the average field of the electrons. Second, the nuclei move on a single adiabatic PES. Third, a PES is approximated by an empirically determined function, called the force field. Last, nuclear motions are described by classical mechanics. Further detailed description of classical MD will be presented in Section 1.4.

In general, the BO approximation can be justified in most physical situations. On the other hand, there are many important chemical phenomena where the BO approximation is invalid, for example, charge transfer and photoisomerization reactions,

which are characterized by the inseparability of electronic and nuclear motion [18], thus factorizing the true wave function in Eq. 1.1 would be clearly a poor approximation. For the systems studied in this thesis, however, the separation of two types of motion is reasonable and so the BO approximation is adopted.

1.3 Statistical Mechanics

Simulation techniques can help us probe the structural properties of molecules and the microscopic interactions between them. This serves as a complement to conventional experiments by providing us with new insights. However, all properties cannot be directly measured in a simulation. Conversely, many quantities that can be obtained in a simulation, do not correspond to properties that are measured in real experiments. Average properties, averaged over a large number of particles, and usually, also averaged over the duration of the measurement are obtained in typical experiments [19]. If we wish to use a simulation technique as a bridge between microscopic length and time scales and the macroscopic world of the laboratory, we must know what kind of averages we should aim to compute. In order to explain this, we need to introduce the language of statistical mechanics.

Statistical mechanics is that branch of physics which studies macroscopic systems from a microscopic or molecular point of view. The goal of statistical mechanics is the understanding and prediction of macroscopic phenomena and the calculation of macroscopic properties from the properties of the individual molecules making up the system [20]. The system could range from a collection of solvent molecules to a solvated protein-DNA complex. In this section, the fundamental concepts of statistical mechanics will be described.

1.3.1 Classical statistical mechanics

One of the most important concepts of statistical mechanics involves asking what is the most probable distribution of energy among a large number of N particles within a container of volume V that is maintained in equilibrium at a specified temperature

T. According to statistical mechanics, the probability P_i of finding the system in its quantum state i with energy E_i in equilibrium at constant N, V and T (the *canonical* ensemble) is proportional to the Boltzmann factor, that is,

$$P_i \propto \exp(-E_i/k_BT) \ , \tag{1.9}$$

where k_B is the Boltzmann constant, and thus we can write the Boltzmann distribution as follows:

$$P_i = \frac{\exp(-E_i/k_BT)}{Q} \ , \tag{1.10}$$

where Q is the partition function, the sum of the Boltzmann factors over all quantum states i:

$$Q = \sum_i \exp(-E_i/k_BT) \ . \tag{1.11}$$

Just as the wave function characterizes the microscopic system in quantum mechanics, the partition function is a fundamental function having an equivalent status in statistical mechanics, and this allows us to calculate many useful temperature-dependent properties of a system such as the internal energy, the heat capacity and the free energy.

The equilibrium value of some observable A is therefore obtained by averaging over all states accessible to the system, weighting each state by the Boltzman factor. Quantum mechanically, this averaging is performed simply by summing over all possible discrete states:

$$\langle A \rangle = \frac{\sum_i \exp(-E_i/k_BT)\langle i|\hat{A}|i\rangle}{Q} \ , \tag{1.12}$$

where $\langle i|\hat{A}|i\rangle$ denotes the expectation value of the operator $\hat{A}$ associated with the observable A in quantum state i. Using the relation of $\exp(-E_i/k_BT) = \langle i|\exp(-\hat{H}/k_BT)|i\rangle$, where $\hat{H}$ is the Hamiltonian of the system, Eq.(1.12) can be written as

$$\langle A \rangle = \frac{\sum_i \langle i|\exp(-\hat{H}/k_BT)\hat{A}|i\rangle}{\sum_i \langle i|\exp(-\hat{H}/k_BT)|i\rangle} \ . \tag{1.13}$$

In a classical statistical mechanical description, the energy is written as a function of positions and momenta, and a state of the system is defined by positions and

momenta of all particles. If the system contains N particles, there are $3N$ positions and $3N$ momenta. It is convenient to visualize the state of an N-particle system in terms of a $6N$-dimensional vector space, called *phase space*. The coordinates are therefore composed of the $3N$ positions $\mathbf{q}$ and $3N$ momenta $\mathbf{p}$. Each point in phase space, called a *phase point*, represents a state of the classical system at any time t. As phase space variables, *i.e.*, ($\mathbf{q}$, $\mathbf{p}$) of the system vary, the point representing the system traces out a trajectory in the phase space. Thus, the corresponding classical expression of a partition function requires an integration over all possible classical "states" of the system instead of a discrete summation because the classical states are continuously defined by all points on the phase space. It is noted that the classical energy is a continuous function of $\mathbf{q}$ and $\mathbf{p}$ for the N particles in the system, *i.e.*, the Hamiltonian function $H(\mathbf{p}, \mathbf{q})$, and now the partition function can be expressed as

$$Q_{cl} = \frac{1}{h^{dN} N!} \int \int d\mathbf{p} d\mathbf{q} \exp(-H(\mathbf{p}, \mathbf{q})/k_B T) \ , \tag{1.14}$$

where h is Planck's constant and d is the dimensionality of the system. The factor $1/N!$ is for taking the indistinguishability of identical particles into account.

Eq. (1.12) for the average value of some observable A can also be simplified to a more workable expression in the classical limit.

$$\langle A \rangle = \frac{\int \int d\mathbf{p} d\mathbf{q} \exp(-H(\mathbf{p}, \mathbf{q})/k_B T) A(\mathbf{p}, \mathbf{q})}{Q_{cl}} \ . \tag{1.15}$$

Eqs. (1.14) and (1.15) are the starting points for virtually all classical simulations of many particle-systems.

1.3.2 Ergodicity

In statistical mechanics, the average value of the observable A is defined as an average over all possible states of the system, called an *ensemble average* denoted by $\langle A \rangle_{ensemble}$. An *ensemble* is the assembly of all possible microstates consistent with the macroscopic constraints on the system. An ensemble average can be written as

$$\langle A \rangle_{ensemble} = \int \int d\mathbf{p} d\mathbf{q} A(\mathbf{p}, \mathbf{q}) \rho(\mathbf{p}, \mathbf{q}) \ , \tag{1.16}$$

where $\rho(\mathbf{p}, \mathbf{q})$ is the probability density of the ensemble defined by the set of variables $(\mathbf{p}, \mathbf{q})$ in turn given by

$$\rho(\mathbf{p}, \mathbf{q}) = \frac{1}{Q_{cl}} \exp(-H(\mathbf{p}, \mathbf{q})/k_B T) \ . \tag{1.17}$$

However, such a phase space average is not the way we usually think about the average behaviour of a system. In most experiments a series of measurements are performed over a long time from which an average is then determined. Another approach used in MD simulations, is to study the average behaviour of a system by computing the natural time evolution of the system numerically and by averaging the quantity of interest over a sufficiently long time. This is called as a *time average.* The time average of A is expressed as

$$\bar{A}_{time} = \lim_{T \to \infty} \frac{1}{T} \int_0^T A(\mathbf{p}(t), \mathbf{q}(t)) dt \ , \tag{1.18}$$

where T is the simulation time. Note that a bar indicates a time average to distinguish it from the above ensemble average, denoted by a $\langle ... \rangle$.

Here, the underlying assumption is called the *ergodic hypothesis*, which states that the time average over a long time is equal to the ensemble average:

$$\bar{A}_{time} = \langle A \rangle_{ensemble} \ . \tag{1.19}$$

Therefore, given enough time, an ergodic system should explore the set of all phase points $(\mathbf{p}, \mathbf{q})$. This is our justification for calculating trajectories (values of $\mathbf{p}(t)$ and $\mathbf{q}(t)$) in order to obtain the thermodynamic properties via time averaging.

Eq. (1.19) tells us that we can use either the time averaging (the Molecular Dynamics (MD) approach) or the ensemble averaging (the Monte Carlo (MC) approach) to calculate the average of a function of the positions and momenta of a many-particle system [19]. The MD and MC methods are the two main simulation techniques in the studies of dynamics in molecular systems. The following section provides a description of MD, which is the major focus of the work presented in this thesis.

1.4 Molecular Dynamics

MD is a central computational tool in statistical mechanics. It computes the equilibrium and transport properties of a classical many-particle system. The word *classical* means that the nuclear motion of the particles follows the laws of classical mechanics. In other words, MD simulates the time evolution of the system based on Newtons second law.

Analogous to traditional experiments, MD simulations can be considered "*in silico*" experiments. Just as in an experiment, we first prepare a sample of the material that we wish to study, prior to the simulation, we select a system and a model describing the relevant chemistry and physics of the constituent particles. The details about the most commonly used models are given in Section 1.5.1. In experimental work, at the data collection stage, we connect the sample to a measuring instrument (*e.g.*, a thermometer), and we measure the property of interest during a certain time interval. We may need to repeat the measurement several times to achieve high accuracy. In MD simulations, Newton's equation $\mathbf{F} = m\mathbf{a}$ is solved for the model system. Previous to the data collection, the system has to reach equilibrium, *i.e.*, the properties of the system no longer change with time. After equilibration, we perform the actual measurement and this is known as production [19].

To start a simulation, initial momenta ($\mathbf{p}$) and positions ($\mathbf{q}$) are assigned to all particles in the system. The initial positions can be obtained from experimental structures, such as the X-ray crystal structure or the solution structure determined by NMR spectroscopy. The velocities, thus momenta, are assigned randomly to each atom from the Maxwell-Boltzmann distribution, a Gaussian function of momenta, at a given temperature. Next, the interactions between constituent atoms need to be described. Having specified the potential energy V, we are now ready to investigate the classical dynamics. The most time-consuming part of almost all MD is the calculation of the force on each atom. The force $\mathbf{F}_i$ acting on an atom i is determined by the gradient of the potential energy. Suppose that we wish to compute the x-component

of the force $\mathbf{F}_i$, that is

$$\mathbf{F}_{x_i} = -\frac{\partial \mathbf{V}}{\partial x_i} \ . \tag{1.20}$$

This force results in an acceleration $\mathbf{a}$ according to Newton's equation of motion $\mathbf{F}_{x_i} = m_i \mathbf{a}_{x_i}$. By knowing the acceleration of each atom in the system, new atomic momenta and atomic positions $(\mathbf{p}(t), \mathbf{q}(t))$ are calculated at regular time intervals, known as the *time step* Δt. Integration of the equations of motion over a long period of time then yields a trajectory that describes the positions and momenta of all particles in the system as they vary with time. The average values of properties can be determined from the trajectory. Several numerical algorithms have been designed for integrating the equations of motion using *finite difference methods.* All algorithms assume that positions, velocities and accelerations can be approximated by a Taylor series expansion. We can derive the simplest but most widely used one, the Verlet algorithm [21] which integrates the equations of motion in an MD simulation. This Verlet algorithm uses the positions r and accelerations a at time t, and the position from the previous step $t - \Delta t$, to calculate the new positions at $t + \Delta t$. The positions at $t + \Delta t$ and $t - \Delta t$ are each expanded as a Taylor series about t:

$$\begin{aligned} r(t + \Delta t) &= r(t) + v(t)\Delta t + \frac{1}{2} a(t) \Delta t^2 \\ r(t - \Delta t) &= r(t) - v(t)\Delta t + \frac{1}{2} a(t) \Delta t^2 \ . \end{aligned} \tag{1.21}$$

Summing these two equations, one obtains the position at $t + \Delta t$

$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) + a(t)\Delta t^2 \ . \tag{1.22}$$

The velocities do not explicitly appear in the Verlet algorithm. The velocities can be calculated by dividing the difference in positions at times $t + \Delta t$ and $t - \Delta t$ by $2\Delta t$:

$$v(t) = \frac{r(t + \Delta t) - r(t - \Delta t)}{2\Delta t} \ . \tag{1.23}$$

It is clear that a good MD program requires a good algorithm to integrate Newton's equations of motion. Some of the features that characterize a 'good' method is being fast, requiring minimal memory, and being easy to program [22]. The conservation of energy in classical mechanics is also an important criterion. Simulations

carried out in the microcanonical, namely constant NVE ensemble (*i.e.*, with fixed number of particles N, volume V and total energy E), require that the total energy be conserved throughout the simulation, *i.e.*, the sum of the kinetic and potential energies is kept constant [23]. Another virtue to be considered is that the algorithm should permit the use of a relatively longer time step Δt. As mentioned earlier, the most demanding part of a MD simulation is the calculation of the forces on each particle in the system. The size of the time step is particularly relevant to the computational demands because the longer the time step, the fewer the number of evaluations of the forces needed per unit of simulation time. Hence, accuracy for long time steps is important and this suggests that it may be advantageous to use a sophisticated algorithm that allows the use of a long time step. A typical time step used in simulations is approximately one order of magnitude smaller than the fastest characteristic time scale of the motion of interest [22]. Lastly, a good algorithm must be reversible in time as in classical mechanics, *i.e.*, if we change the signs of all velocities, a trajectory retraces itself backward in time.

Constraint algorithms are often applied to MD simulations. A constraint algorithm is a method for satisfying constraints for molecular systems that obey Newton's equations of motion. The SHAKE method [24] is the first algorithm developed to satisfy bond geometry constraints with the use of Verlet algorithm. This can enable the time step in a MD simulation to be increased. For the studies reported in this thesis, the SHAKE algorithm was used for constraining bonds involving hydrogen atoms due to their much higher vibrational frequencies.

MD has been widely applied and has brought and/or could bring important contributions in areas such as liquids, clusters, surfaces, biomolecules, and so on. In particular, a classical description of the nuclear dynamics allows us to study the dynamics of large macromolecules, including biological systems such as proteins, nucleic acids (DNA, RNA), and membranes [25]. Dynamical events may play a key role in controlling processes which affect functional properties of biomolecules. MD simulations are commonly used to design drugs in the pharmaceutical industry to test properties virtually without actual synthesis. It is also worth mentioning that the

development of *ab initio* MD such as the Car-Parrinello method [26] where the forces on atoms are obtained by solving for the electronic structure at each time step not through nuclear potential, allows us to study simultaneously electronic as well as dynamical properties.

1.4.1 Constant temperature molecular dynamics

In a conventional MD simulation, the total energy E is a constant of motion. Hence, MD simulations measure time averages in the NVE ensemble. However, we often encounter limitations and inconveniences which come from the use of the NVE ensemble. Many laboratory experiments are carried out at constant temperature T and constant pressure P while MD simulations are performed at constant energy E and constant volume V. The difference in these conditions makes direct comparison with experiments difficult [27]. Although thermodynamic results can be transformed between ensembles, this is strictly only possible in the limit of infinite system size ("the thermodynamic limit"). Therefore, it is often desirable to perform simulations in other ensembles, commonly the canonical (constant temperature NVT) ensemble or the isothermal-isobaric (constant pressure NPT) ensemble. In order to maintain the temperature constant on average, some modifications are applied to the standard MD algorithm discussed earlier, and this is called a *thermostat* algorithm. In this section, several such methods will be discussed.

The temperature of the system is related to the ensemble average of the kinetic energies of all particles:

$$\langle K \rangle_{NVT} = \langle \sum_i \frac{1}{2} m_i v_i^2 \rangle = \frac{3}{2} N k_B T \ , \tag{1.24}$$

where v_i is the velocity of particle i and N is the total number of particles in the system.

There have been various types of different thermostats to simulate the constant temperature condition. Since the temperature depends on the velocities, the crude and simplest way to control the temperature is to multiply the velocities at each time

step by the factor $\lambda = \sqrt{T_w/T(t)}$, where T_w is the desired temperature and $T(t)$ is the current temperature at time t [22].

An alternative way to maintain the temperature is to couple the system to an external heat bath that is fixed at the desired temperature [28]. The bath acts as a source of thermal energy, adding or removing heat from the system as appropriate. This method, called Berendsen thermostat, "encourages" the temperature in the desired direction by coupling it to a heat bath. It corrects deviations of the temperature $T(t)$ from the desired temperature T_w by scaling the velocities at each time step Δt (and hence control the value of temperature). The scaling factor for the velocities is:

$$\chi^2 = \left(1 + \frac{\Delta t}{\tau}\right)\left(\frac{T_w}{T(t)} - 1\right), \tag{1.25}$$

where τ is the coupling parameter.

The Langevin thermostats follow the Langevin equation of motion instead of Newton's equation of motion [29]. In the Langevin equation of motion, a frictional force, proportional to the atomic velocity, is added to the conservative force with the purpose of adjusting the kinetic energy of the particles such that the temperature matches the desired temperature.

There are also other methods to maintain constant temperature in MD: stochastic (*e.g.*, Andersen thermostat [30, 27]) and extended system methods (*e.g.*, the Nosé [31] and Nosé-Hoover thermostat [32, 33]). For the studies reported in this thesis, the Berendsen and Langevin thermostats were used, and simulations were performed in the NPT and NVT ensembles.

1.4.2 Equilibrium properties

Properties of interest of many-body systems can be "measured" from molecular simulations. Such properties include, of course, those quantities that can be compared with real experimental observables for the thermodynamic properties of the system under consideration. Examples are the temperature T, pressure P, and heat capacity C_v. As an illustration, consider the temperature. Section 1.4.1 gives the relationship between the temperature and the average kinetic energy. For a system with $3N$

degrees of freedom, the temperature can be estimated as

$$T = \frac{2\langle K \rangle}{3Nk_B} \,. \tag{1.26}$$

In MD simulations, it is a common practice to impose geometrical constraints on certain high frequency motions, *e.g.*, hydrogen-containing bonds. In such cases, the effective number of degrees of freedom is $3N - N_c$ where N_c is number of constraints.

However, some thermodynamic functions cannot be obtained directly in a simulation. In other words, these properties cannot be expressed as a simple average of the phase space coordinates of all the particles in the system [19]. Examples of such properties are the entropy S, the Helmholtz free energy A, and the Gibbs free energy G. We require separate techniques to evaluate such quantities in computer simulations. Methods used to calculate these properties will be described in Section 1.4.4.

Another property that can be obtained from simulation is the so-called *radial distribution function* (also known as the *pair correlation function*), denoted by $g(r)$. This function is a useful tool to describe the structure of a system, particularly of liquids, and can be easily obtained in a simulation. In particular, $g(r)$ is defined simply as the ratio between the average number density $\rho(r)$ at a distance r from any given atom and the density at a distance r from an atom in an ideal gas at the same overall density. By construction, $g(r) = 1$ for an ideal gas. Any deviation of $g(r)$ from unity reflects correlations between the particles due to the intermolecular interactions [19, 34].

Both thermodynamic and structural properties do not depend on the time evolution of the system. They are static equilibrium averages (time-independent). Such equilibrium quantities can be obtained by either MD or MC simulations.

1.4.3 Dynamical properties

In addition to the static equilibrium properties, dynamical properties (time-dependent) can be obtained in an MD simulation, but not in an MC simulation. This is a major advantage of the MD over the MC method.

Onsager's regression hypothesis [35, 34] states that the law governing the regression of microscopic thermal fluctuations in a system at equilibrium is identical to the law describing the relaxation of a macroscopic system in a state away from equilibrium, provided the perturbation to the system is very weak. In other words, by observing (microscopic) fluctuations in simulations of an equilibrated system, one can learn about its (macroscopic) dynamical properties.

In order to understand this hypothesis, consider an observable A for a system at thermal equilibrium. Such a property fluctuates in time with the following *spontaneous microscopic fluctuations*,

$$\delta A(t) = A(t) - \langle A \rangle \ , \tag{1.27}$$

where $A(t)$ is the instantaneous value of the observable A at time t and $\langle A \rangle$ is the corresponding equilibrium ensemble-averaged quantity. The time evolution of the fluctuation is governed by microscopic laws. For classical systems, it really is a function of time with parametric dependence on positions $\mathbf{q}$ and momenta $\mathbf{p}$:

$$\delta A(t) = \delta A(t; \mathbf{p}, \mathbf{q}) = \delta A(\mathbf{p}(t), \mathbf{q}(t)) \ . \tag{1.28}$$

The average correlation between $\delta A(t)$ and an instantaneous fluctuation at time zero $\delta A(0)$ is described by the *correlation function* [34]

$$C(t) = \langle \delta A(t) \delta A(0) \rangle = \langle A(t) A(0) \rangle - \langle A \rangle^2 \ . \tag{1.29}$$

Thus, for a classical system, the time correlation function of A is defined as

$$C(t) = \int d\mathbf{p} \int d\mathbf{q} \ f(\mathbf{p}, \mathbf{q}) \delta A(0; \mathbf{p}, \mathbf{q}) \delta A(t; \mathbf{p}, \mathbf{q}) \ , \tag{1.30}$$

where $f(\mathbf{p}, \mathbf{q})$ is the equilibrium phase space distribution function.

In equilibrium statistical mechanics, all thermodynamic properties can be accessed through the partition function of the system under investigation. Similarly, time correlation functions are central to the calculation of transport properties [36]. There is however one important difference: while the state of thermal equilibrium is uniquely defined, there exists a wide collection of different nonequilibrium states. In practice,

this means that *all* thermodynamic quantities can be obtained from *one* partition function. Conversely, a specific time correlation function is required for *every* different transport process.

The below is an example of the connection between a time correlation function and transport properties [34]:

$$D = \frac{1}{3}\int_0^\infty \langle \mathbf{v}(t) \cdot \mathbf{v}(0) \rangle dt \ . \tag{1.31}$$

This equation relates the diffusion constant D to an infinite time integral of the velocity $\mathbf{v}$ autocorrelation function. In the present, we will specifically focus on the hydrogen bond lifetime correlation function that will be defined in Chapter 5.

1.4.4 Free energy simulations

The free energy is arguably the most useful quantity in thermodynamics. It is a state function of a system in thermodynamic equilibrium, often expressed as the Helmholtz function, or the Gibbs function. Whether one chooses to describe the free energy in terms of the Helmholtz or Gibbs function depends on the system under investigation. The Helmholtz free energy A provides a suitable description for systems specified by a fixed number of particles and constant temperature and volume (NVT ensemble) while the Gibbs free energy G is more appropriate for an ensemble with constant number of particles, temperature and pressure (constant NPT). Most experiments are conducted under conditions of constant temperature and pressure so we usually require the Gibbs free energy for proper comparison with experimental values.

In classical statistical mechanics, A and G are directly related to the canonical (N, V, T) partition function and (N, P, T) partition function, respectively.

$$\begin{aligned} A &= -k_B T \ln Q(N, V, T) \\ G &= -k_B T \ln Q(N, P, T) \ . \end{aligned} \tag{1.32}$$

Unfortunately, it is very difficult to compute the absolute free energies because this quantity depends on the absolute value of the partition function, which requires

an extensive sampling of phase space. Thus, associated quantities such as the entropy and the chemical potential are also difficult to calculate. As we mentioned in Section 1.4.2, the free energy cannot be accurately determined from a 'standard' MD or MC simulation. It is, however, simpler to compute relative free energies, which depend on a much smaller region of phase space. Three methods have been proposed for calculating free energy differences: thermodynamic perturbation, thermodynamic integration and slow growth [22, 37]. The strategy employed in these methods relies on chemical 'mutations'. A thermodynamic integration is often used to calculate the difference in excess free energy of similar but distinct molecules, and such calculations are particularly important in biomolecular modelling [38]. For the studies in this thesis, we will not focus on these methods to calculate free energies.

Instead, we look at the change in free energy as a function of some inter- or intramolecular coordinates, such as the distance between two atoms, or the torsion angle of a bond within a molecule, or in other words, the variation of free energy along conformational degrees of freedom. The free energy surface along a chosen coordinate is known as a *potential of mean force* (PMF). The simplest type of PMF is the change in free energy as two tagged particles are moved through the system from infinite separation to a relative separation r. We can calculate the PMF from the radial distribution function using the following expression for the Helmholtz free energy [34]:

$$w(r) = -k_B T \ln g(r) \; . \tag{1.33}$$

A detailed description of the PMF will be provided in the next section.

1.4.5 Umbrella sampling

The PMF $w(\chi)$ along some coordinate χ was first introduced by Kirkwood [39] in 1935. It is a key concept in modern statistical mechanical theories of liquids and of complex molecular systems. The PMF is defined in terms of the average distribution function $\langle \rho(\chi) \rangle$ as

$$w(\chi) = w(\chi^*) - k_B T \ln \left[\frac{\langle \rho(\chi) \rangle}{\langle \rho(\chi^*) \rangle} \right] , \tag{1.34}$$

where $w(\chi^*)$ and $\langle\rho(\chi^*)\rangle$ are arbitrary reference values. The average distribution function along the coordinate is obtained from the average weighted by the Boltzmann factor:

$$\langle\rho(\chi)\rangle = \frac{\int d\mathbf{q}\delta(\chi'(\mathbf{q}) - \chi)\exp(-V(\mathbf{q})/k_BT)}{\int d\mathbf{q}\exp(-V(\mathbf{q})/k_BT)} , \tag{1.35}$$

where $\delta(\chi'(\mathbf{q}) - \chi)$ is the Dirac function for the coordinate χ and $\chi'(\mathbf{q})$ is a function depending on a few or several degrees of freedom. The chosen coordinate χ is assumed as a geometrical coordinate $\chi(\mathbf{q})$.

Unlike the frequently used mutations of free energy perturbation calculations [22, 37, 40] which are often along non-physical pathways, the PMF is usually calculated for a physically achievable process. In particular, it is useful for calculating conformational equilibrium properties and for predicting the transition rate of dynamically activated processes.

However, it is often impractical to compute the PMF $w(\chi)$ or the distribution function $\langle\rho(\chi)\rangle$ directly from MD simulations. For example, systems in which a large potential energy barrier separates two regions of configurational space may suffer from poor sampling in a simulation. In other words, the low probability of overcoming the potential barrier can leave inaccessible configurations on the other side of the barrier poorly sampled or even entirely unsampled within the available computer time. To avoid this problem, special sampling techniques have been designed to calculate the PMF. One of these approaches is called umbrella sampling [41]. In umbrella sampling, a modified-Boltzmann sampling scheme is used to avoid sampling difficulties by modifying the potential function. The modification of the potential function can be written by adding artificial biasing potentials V_b to the potential energy $V(\mathbf{q})$:

$$V'(\mathbf{q}) = V(\mathbf{q}) + V_b(\chi) . \tag{1.36}$$

This forces the system to compute an ensemble average over a modified-Boltzmann distribution within a small interval of a prescribed value of χ. Biasing potentials are added over a range of coordinates, called the *window*. Multiple simulations (windows) are performed with different biasing potentials $V_b(\chi)_i$, centred on successive values of

χ for every window i. The biasing potential is often expressed as a harmonic function of the form:

$$V_b(\chi)_i = \frac{1}{2}k(\chi - \chi_i)^2 \ , \tag{1.37}$$

where is k is the harmonic force constant. At every window i, the biased distribution function $\langle\rho(\chi)\rangle_i^b$ is then obtained through the expression of the distribution function in Eq. (1.35) substituting $V(\mathbf{q})$ for $V'(\mathbf{q})$.

$$\langle\rho(\chi)\rangle_i^b = \exp(-V_b(\chi)_i/k_BT)\langle\rho(\chi)\rangle_i^u\langle\exp(-V_b(\chi)_i/k_BT)\rangle^{-1} \ . \tag{1.38}$$

The superscripts b and u indicate *biased* and *unbiased* respectively. Since the window potential $V_b(\chi)_i$ is a known function, we can calculate the unbiased PMF for each ith window from the biased distribution:

$$w(\chi)_i^u = w(\chi^*) - k_BT\ln\left[\frac{\langle\rho(\chi)\rangle_i^b}{\langle\rho(\chi^*)\rangle}\right] - V_b(\chi)_i + F_i \ , \tag{1.39}$$

where F_i are undetermined constants that represent the free energy associated with introducing the window potential. Within each window we get a free energy profile with a different constant F_i. The constants F_i are obtained by adjusting the various adjacent windows $w(\chi)_i^u$ in the region in which they overlap until they match.

The distribution functions from various windows need to be unbiased $\langle\rho(\chi)\rangle_i^u$ (the modified-Boltzmann factor is removed) and then recombined together to obtain the final estimated PMF $w(\chi)$. The process of unbiasing and recombining the different simulation windows is the main difficulty in umbrella sampling. The different methods available to unbias and recombine data extracted from umbrella sampling was reviewed by Roux [41]. One useful method for calculating free energies is called the Weighted Histogram Analysis Method (WHAM) [42]. WHAM represents a generalization and an extension of the histogram developed by Ferrenberg and Swendsen [43]. It allows us to obtain better estimates by combining the results of all the different simulations. The WHAM equation expresses the total unbiased distribution function $\langle\rho(\chi)\rangle^u$ as a χ-dependent weighted sum over the N_w individual unbiased distribution functions $\langle\rho(\chi)\rangle_i^u$,

$$\langle \rho(\chi) \rangle^u = \sum_{i=1}^{N_w} \langle \rho(\chi) \rangle_i^u \times \left[\frac{n_i \exp(-[V_b(\chi)_i - F_i]/k_B T)}{\sum_{i=1}^{N_w} n_j \exp(-[V_b(\chi)_j - F_j]/k_B T)} \right] , \tag{1.40}$$

where N_w is the number of windows and n_i is the number of independent data points used to construct the biased distribution function. This equation can also be expressed in terms of known biased distribution functions $\langle \rho(\chi) \rangle_i^b$.

$$\langle \rho(\chi) \rangle^u = \sum_{i=1}^{N_w} n_i \langle \rho(\chi) \rangle_i^b \times \left[\sum_{j=1}^{N_w} n_j \exp(-[V_b(\chi)_j - F_j]/k_B T) \right] . \tag{1.41}$$

One of the main advantages of the WHAM method is that it can be easily extended to treat the case of a PMF depending on more than one variable.

1.5 Computer Modelling

Computer simulations have become a useful part of mathematical modelling of many natural systems in physics, chemistry and biology [44]. Results presented in the current thesis are obtained from molecular dynamics simulations, a form of computer simulation described earlier in Section 1.4.

Simulations make use of a model to represent real world phenomena or objects. This representation often takes the form of mathematical equations. Traditionally, mathematical models attempt to find analytical solutions to problems and enable us to predict the behaviour of the system from a set of parameters and initial conditions.

Modelling is an important field of computational research in general. The accuracy of simulated results is affected by the choice of model. Usually there is no single ideal model capable of treating all problems. Often a model will give an accurate reproduction or prediction of experimental measurements for certain compounds and fail miserably for others. Therefore, the model should be chosen carefully for a given problem.

In theoretical and computational chemistry specifically, molecular modelling attempts to predict physical properties for molecular systems based on the numerical solution to the equations that embody the physical laws governing the behaviour.

At the most fundamental level, this approach involves the direct solution to the Schrödinger equation for the nuclear and electronic degrees of freedom. In practice, several approximations are used. For instance, in MD simulations, a classical description of nuclear motion is employed and electronic effects are taken into account in the model. A general introduction to the molecular model used in this thesis is given below.

The ultimate aim in molecular modelling is both to explain experimental observations and to act in a predictive capacity. Some problems are not amenable to experiments or some questions are rather hypothetical in nature. In contrast, a model can be easily built and studied, and problems encountered during experiments can be within the capabilities of modelling. Thus, a good modelling study allows us to provide a framework for integrating the experimental results from various techniques in order to provide a greater overall understanding of the problem of interest.

1.5.1 Molecular modelling

For any calculation in science a model must be constructed, which must in some way approximate reality. The underlying model for classical MD simulations, is that the energy of a molecule can be described in terms of a function called the *force field*. Force field refers to the functional form and parameter sets used to describe the potential energy of a system. As mentioned earlier in Section 1.2, this is a fully classical potential and the electrons are considered implicitly in terms of the potential energy surface on which the atoms move.

The basic functional form of a force field consists of both bonded terms relating to atoms that are linked by covalent bonds, and non-bonded terms describing the long-range electrostatic and van der Waals forces. A typical form for the total energy in an additive force field can be written as $E_{tot} = E_{bonded} + E_{nonbonded}$ where the components of the bonded contributions are given by

$$E_{bonded} = E_{stretch} + E_{bend} + E_{torsion} \ . \tag{1.42}$$

$E_{stretch}$ and E_{bend}, representing the bond stretching and bond angle bending respec-

tively are usually modelled as harmonic oscillators:

$$\begin{aligned} E_{stretch} &= \frac{1}{2}k_r(r - r_0)^2 \\ E_{bend} &= \frac{1}{2}k_\theta(\theta - \theta_0)^2 \ , \end{aligned} \tag{1.43}$$

where k_r and k_θ are harmonic force constants for bond stretching and bond angle bending, respectively. The r and θ are the actual bond length and bond angle and r_0 and θ_0 are their corresponding values at equilibrium. The torsional angle interaction given by $E_{torsion}$ in Eq. 1.42 is expressed as

$$E_{torsion} = \sum_n \frac{1}{2}V_n(1 + \cos(n\tau + \delta_n)) \ , \tag{1.44}$$

where V_n and τ are the torsional rotation force constants and the current torsional angle respectively. δ_n is the phase angle and the n parameter controls the periodicity.

Non-bonded interactions are usually divided into two:

$$E_{nonbonded} = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{r_{ij}} + 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \ , \tag{1.45}$$

where ϵ_0 is the permittivity of vacuum, q_i and q_j are the charges on atoms i and j respectively, and r_{ij} is the distance between i and j. The quantity ϵ_{ij} is the Van der Waals well depth and σ_{ij} is the distance at which $E_{vdw}^{LJ} = 0$. The first term on the right-hand side corresponds to electrostatic interactions arising from the unequal distribution of charges in a molecule. Within most current force fields, this uneven distribution of charge can be modelled by placing point charges at each atom (*i.e.*, approximating the charge distribution as a single point bearing the charge). The interaction between these point charges is generally modelled by a Coulomb potential, given by the first term. The second term, called the Lennard-Jones potential is one of the most common ways of expressing the van der Waals interactions.

Once a particular form for a force field has been chosen, a set of parameters for each type of atom has to be determined. The typical parameter set includes values for atomic mass, van der Waals radius, and partial charge for individual atoms, and equilibrium values of bond lengths, bond angles, and dihedral angles for pairs,

and values corresponding to the effective spring constant for each potential. Force field parameters are usually determined from either experimental data or electronic structure calculations.

There exists a variety of force fields which use different forms for the various interactions within and between molecules. The particular form of a force field depends on the accuracy required for its intended purpose. For example, the force fields MM2 (Molecular Mechanics 2) [45] and MM3 [46] were developed primarily for conformational analysis of small organic molecules with the aim of making accurate predictions of molecular structures and properties. A number of force fields have been developed for application to biologically interesting molecules such as proteins, nucleic acids or polymers. Typical force fields in this category are the AMBER (Assisted Model Building and Energy Refinement) [47], CHARMM (Chemistry at HARvard Macromolecular Mechanics) [48], OPLS (Optimized Potential for Liquid Simulations) [49] and GROMOS (GROningen MOlecular Simulation) [50] force fields. AMBER is widely used for protein and DNA, and this force field was used for the studies presented in this thesis.

The modelling of carbohydrates and protein-carbohydrate complexes is discussed separately in the following sections. The parameterization of carbohydrates is addressed in Chapter 2. In particular, our work on the charge derivation to model five-membered ring molecules in solution is presented. The RESP approach [51], an electrostatic potential based method using charge restraints, is used to obtain partial atomic charges for furanose ring molecules.

1.5.2 Modelling of carbohydrates

Carbohydrates are one of four major classes of macromolecules in biology (DNA, proteins, carbohydrates, and lipids). They are fundamentally involved in important biological phenomena, such as antibody-antigen interactions and bacterial infection. It is therefore essential to correctly understand the spatial and dynamic properties of carbohydrates [4, 52]. This requires experimental techniques that characterize the structure and dynamics of carbohydrates.

In contrast to proteins which tend to hold their globular shapes, carbohydrates are bendy and twisty, and sometimes they have highly branched strands that are bigger than proteins. Proteins only have a single type of linkage, *i.e.*, amide bonds between monomeric units. Carbohydrates, on the other hand, can be connected to one another via various linkages [4]. The glycosidic oxygens that link together monosaccharides are the place that gives rise to the flexibility of carbohydrates. That floppiness makes carbohydrates difficult to crystallize, so their structures are frequently elucidated by NMR, rather than the X-ray crystallography that is traditionally used for proteins. NMR structures represent averages, however, and it is difficult to say which groups of conformers play important roles in biological processes. Computer modelling helps to narrow down the shapes that carbohydrates are likely to adopt in a given environment, and thus complement experimental data. Computational methods, such as MC and MD simulations, are employed increasingly to augment the experimental approaches in studying the structural and conformational behaviour of carbohydrates.

The dramatic increase in computing power, speed and improved software have led to significant progress in modelling. Numerous force fields and parameter sets have been derived for carbohydrates: HSEA (Hard Sphere Exo-Anomeric) [53], MM2 [54], MM3 [46, 55], CHARMM [48], AMBER [56] and GROMOS [50]. However, very few have gained widespread recognition for application to MD simulations of oligosaccharides. This is due, in part, to a lack of suitable experimental data to use as benchmarks in the testing of the force fields [57] and the lack of accurate torsional energy profiles [12]. Many researchers who are interested in carbohydrate modelling have developed a suitable set of carbohydrate parameters. The GLYCAM (Oligosaccharide/glycoprotein force field) parameter sets developed by Woods and co-workers [12] has been of great practical use in working with oligosaccharides [58, 59] and protein-carbohydrate complexes [60]. For example, one study showed that the parameters allowed the correct prediction of the subtle effects on the rotational properties of the glycosidic linkages in models of methyl α-D-gluco- and α-D-mannopyranoside [12]. The strength of this parameter set comes in large part from its very careful treatment of electrostatic interactions. In GLYCAM, unique partial atomic charges for each

atom within a sugar unit were computed, whereas the majority of other carbohydrates parameter sets assume that certain atom types will be equivalent and transferable between sugars. This first version of GLYCAM (GLYCAM_93) was designed with the intention that it would introduce the minimal parameters necessary to add carbohydrate functionality to the AMBER force field for proteins and nucleic acids. Furthermore, explicit solvent models for water and several other small molecules are available for use with AMBER. Thus, this makes the combination of GLYCAM and AMBER a powerful tool for modelling protein-carbohydrate complexes. Recently, the new derivations of a highly consistent and transferable parameter set for modelling carbohydrates and glycoconjugates (GLYCAM04, GLYCAM06) were reported [11]. This new parameter set removed its previous specificity for carbohydrates and its dependency on the AMBER force field and parameters. GLYCAM_93 and GLYCAM04 were used for the studies in this thesis.

1.5.3 Modelling of protein-carbohydrate interactions

Interactions between proteins and carbohydrates are also amenable to computational approaches. Since protein-carbohydrate interactions play a critical role in many biological processes, a thorough understanding of both carbohydrate and protein structure is essential to predict these interactions. The computational methodology to model carbohydrate-protein complexes has been developed to accurately predict structures of protein-carbohydrate complexes and better understand the details of the interactions at the atomic level [61, 52, 57].

However, modelling protein-carbohydrate complexes is complicated compared to modelling other small molecules because the inherent flexibility of carbohydrates and water-mediated hydrogen bonds to proteins make the simulation of the complexes difficult. Additional difficulties in predicting structures of particular protein-carbohydrate complexes come from the dynamic nature of the structure, allowing the carbohydrate to bind to the protein in multiple conformations [62]. Moreover, a theoretical quantification of the energies involving protein-carbohydrate interactions is more difficult. It is difficult to predict the free energy of binding of a protein-

carbohydrate complex because the relative contribution of enthalpic and entropic terms to the free energy of formation of these complexes varies from receptor to receptor, as well as for different ligands binding to the same receptor [57]. Also, a considerable solvent contribution to the energies associated with protein-carbohydrate interactions makes the prediction of the energies difficult [63, 64]. Consequently, a good model (*i.e.*, an accurately parameterized forcefield) and consideration of solvent effects are necessitated. More sophisticated models for calculating interaction energies of protein-carbohydrate complexes have been used in free energy simulations using explicit solvent [60].

Several recent reviews have indicated that many advances have been made both in methodology and in approaches to validation to model carbohydrates and protein-carbohydrate interactions [65, 52, 66, 67]. Despite this progress, currently available methods are still limited in their ability to describe certain aspects of real system such as proton exchange and anomerization. Nevertheless, modern computational approaches can provide insight into physical properties, some of which are not accessible experimentally.

1.5.4 Overview of thesis

To reiterate, our goal is to develop tools to contribute to carbohydrates modelling, specifically, understanding the solution conformations of carbohydrates containing furanose rings and to elucidate the microscopic details of intrinsic binding interactions between proteins and carbohydrates.

In Chapter 2, we present an approach that we developed for the derivation of charges to model furanosides in solution. This approach was first tested on a monosaccharide, the methyl-α-D-arabinofuranoside. Following this development, a model used for its description was validated by comparison with NMR experiments. Chapter 3 involves the elucidation of the intermolecular interactions within desolvated protein-trisaccharide ligand complexes. This work relies on the interplay between theory and experiment. Simulation results are validated using mass spectrometry combined with electrospray ionization. In Chapter 4, we extend our study to investigate the dy-

namics of water molecules in charged and hydrated protein-trisaccharide complexes in the gas phase. Simulations are performed over a wide temperature range to study hydrogen bond dynamics in gas phase complexes with the aim of probing the fate of individual water molecules. In Chapter 5, theoretical approaches for the prediction of binding constants of protein-trisaccharide complexes are developed and such an approach allows the direct calculation of dissociation rate constants in the context of transition state theory. This type of analysis requires the construction of the potential of mean force along the dissociation reaction coordinate. Concluding remarks and future work are presented in Chapter 6.

Chapter 2

New Model for Furanose Rings

2.1 Introduction

Carbohydrates are involved in important biological functions as mentioned in Section 1.5.2. Thus, a complete knowledge of the conformational properties of carbohydrates is essential to understand their mechanisms of action, which may aid in the design of carbohydrate-based vaccines, and other therapeutic agents. For these reasons, studies of the three-dimensional structures and dynamics of oligosaccharides and polysaccharides have been extensively performed [66, 68, 69, 70, 71, 72, 73, 74].

Unlike polypeptides and proteins, oligosaccharides bend and twist. They do not form well-organized tertiary structures in solution. Rather, oligosaccharides often populate multiple conformational families, thus requiring correct determination of their spatial and dynamic properties. Experimental structure determination methods such as X-ray crystallography and NMR spectroscopy have been applied in studies of carbohydrate conformation. X-ray crystallography generally results in a single

three-dimensional structure of the oligosaccharides [8], which is due to a direct result of difficulties in crystallizing the generally flexible oligosaccharides. This fails to sufficiently describe its dynamic properties. NMR spectroscopy is a more preferred technique to conduct solution conformational analysis. However, the conformations of the glycosidic linkages in these flexible systems are particularly difficult to determine by NMR spectroscopy [11, 75].

Theoretical methods, such as MD simulations, are increasingly applied in determining the conformational properties of oligosaccharides. MD simulations offer the advantage of providing valuable information complementary to the experimental results. Numerous force fields and parameter sets have been derived for carbohydrates [76]. However, over the past several years the use of the AMBER force field [77] in conjunction with the GLYCAM carbohydrate parameter set [12, 11] has emerged as a reliable force field in working with oligosaccharides containing six-membered pyranose ring forms [58, 59]. For example, the conformation of hydroxymethyl groups on pyranosides has been studied by Kirschner and Woods [68]. In this study, quantum mechanics calculations and solvated MD simulations were performed on two representative carbohydrates, methyl α-D-glucopyranoside and methyl α-D-galactopyranoside. It showed that correct reproduction of the experimental rotamer populations about the ω-angle ($O_6 - C_6 - C_5 - O_5$) was obtained only after explicit water was included in the MD simulations. It also provided a quantitative explanation of the conformational behaviour of oligosaccharides containing glycosidic linkages at the 6-position ($1 \rightarrow 6$ linked).

We have performed a similar conformational study of hydroxymethyl groups on five-membered *furanoside* rings. MD simulations were carried out on methyl-α-D-arabinofuranoside (**1**) using the AMBER force field and the GLYCAM carbohydrate parameter set. The notable differences from the studies on the conformation of pyranoside rings [68] are that ring conformations are addressed and a novel charge derivation approach that accounts for the flexibility of the furanoside ring is proposed. The following sections will focus on the theoretical modeling of furanosides, including a description of the new charge derivation procedure.

2.2 Furanose Ring Systems

Furanose rings are important components of a number of glycoconjugates: nucleic acids [78], bacterial, parasitic, and fungal cell wall polysaccharides [79, 80], as well as other natural products [81, 82]. Two examples of these glycoconjugates are arabinogalactan (AG) and lipoarabinomannan (LAM), which are present in the cell wall of *Mycobacterium tuberculosis*, the organism that causes tuberculosis, and other mycobacteria [83]. The AG, a polysaccharide containing approximately 100 monosaccharide units, is composed entirely of arabinofuranose and galactofuranose residues, except for two pyranose moieties, which serve as the linker between the glycan and peptidoglycan [84]. Similarly, a significant component of LAM is an arabinan domain, representing approximately half the molecular weight, which contains only arabinofuranose residues [85].

Recent progress in the field of carbohydrate chemistry has provided an understanding of the three-dimensional structure of oligosaccharides containing pyranose rings by X-ray crystallography and NMR spectroscopy [86]. On the other hand, the conformational preferences of oligosaccharides composed of furanose rings are not well understood. This is due to a lack of experimental data on oligofuranosides and the inherent flexibility of five-membered rings, which profoundly influences their role in biological processes. Consequently, a greater understanding of the conformational preferences of these ring systems is an important area of research.

Early studies in the area of furanoside conformation focused on defining the conformational preferences of the sugar residues present in the nucleic acids: D-ribofuranose in RNA and 2-deoxy-D-*erthyro*-pentofuranose (2-deoxy-D-ribose) in DNA [78]. A model developed by Altona and Sundaralingam through analysis of a large number of nucleoside crystallographic structures can be used to describe conformational preferences of any furanose ring [87, 88]. This model makes use of the pseudorotational itinerary (see Figure 2.1) to describe the possible ring conformers. Conformationally, furanose rings can adopt a number of envelope (E) and twist (T) conformers that are separated by typically low-energy barriers. Each conformer is described by two

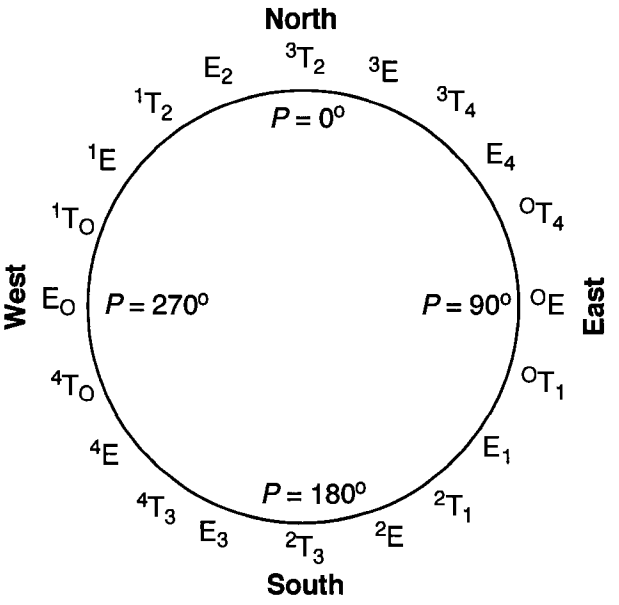


Figure 2.1: Pseudorotational itinerary for a D-aldofuranose ring.

parameters, the pseudorotational phase angle (P) and the puckering amplitude (ϕ_m), which can be calculated from five endocyclic torsion angles (see Figure 2.2) of the ring [89]:

$$\begin{aligned} \tan P &= \frac{(\phi_2 + \phi_4) - (\phi_1 + \phi_3)}{3.077\phi_0}, \\ \phi_m &= \frac{\phi_0}{\cos P}. \end{aligned} \tag{2.1}$$

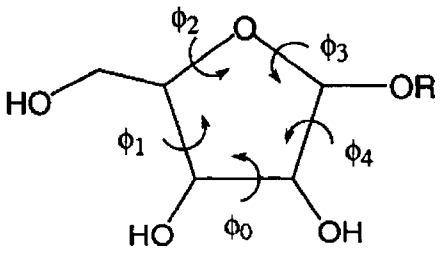


Figure 2.2: Definition of endocyclic torsion angles ϕ_0 - ϕ_4.

In solution, there is a dynamic equilibrium between two ring conformers, the North (N) and South (S) conformers, found in the northern hemisphere and southern hemisphere of the pseudorotational itinerary, respectively [87]. Conformational investigations of furanoside rings by NMR spectroscopy most commonly involve analysis using PSEUROT [90], a program that assumes this two-state equilibrium and which fits the experimental ^{1}H-^{1}H coupling constant data to two conformers and their populations.

Due to the inherent flexibility of furanosides, their conformational analysis is much more complicated than similar studies with pyranosides. In addition to the torsional flexibility of the ring, other key conformational features are of importance. These include rotamer populations about the glycosidic C1-O1 and C4-C5 bonds. The preferred rotamer about the C1-O1 bond places the aglycone (*e.g.*, the methyl group in furanoside) *anti* to the C1-C2 bond, as this is favoured by the *exo*-anomeric effect [91]. For the C4-C5 bond (ω angle), three rotamers are typically present, *gt*, *tg* and *gg* (see Figure 2.3), with the distribution being influenced by a combination of steric and stereoelectronic (*gauche*) effects [92, 93, 94, 95].

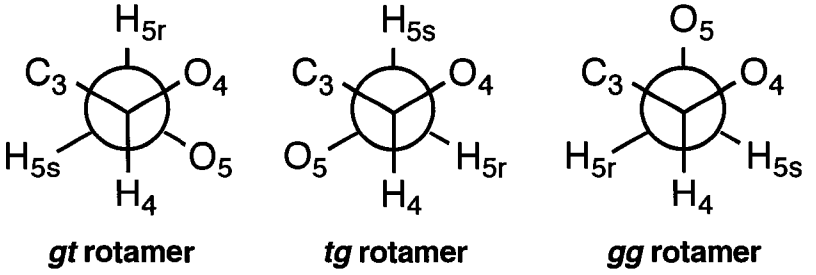
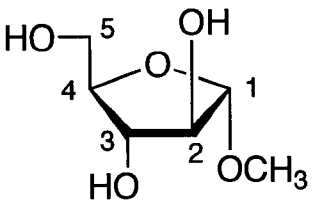


Figure 2.3: Definition of *gt*, *tg* and *gg* rotamers about the C4-C5 bond.

Over the past years, a series of NMR studies on the arabinofuranose-containing oligosaccharides [96, 89] were carried out and these experimental studies were coupled with high-level *ab initio* and density functional theory calculations on methyl α-D-arabinofuranoside (**1**, see Figure 2.4) [97, 98] and related analogs [99, 100, 101, 102].



1

Figure 2.4: Structure of methyl α-D-arabinofuranoside.

Having studied the conformation of **1** using both experimental and high-level computational methods, we are interested in looking at larger oligomers of D-arabinofuranose, for which we have NMR data [96, 89]. However, given the size of these molecules, their treatment with *ab initio* or density functional theory methods is of limited practicality. Thus, we have begun to investigate the use of force field models to probe the conformation of these oligosaccharides. Previous molecular mechanics studies of furanosyl rings have largely been carried out using MM3 or earlier variants of this forcefield [103, 104, 105, 106, 107, 108, 67]. The GLYCAM parameter set for carbohydrates [12] has been of great practical use in working with oligosaccharides containing pyranose rings [58, 59]. Here, we describe the results of our first investigations of the use of the GLYCAM parameters and the AMBER force field to study the conformation of furanoside rings. We initially studied the ability of this computational method to predict the rotamer distribution about the C4-C5 bond and pseudorotational phase angle in **1** as determined by NMR spectroscopy. We also propose a new charge derivation approach to consider the flexibility of the furanoside ring by taking an average of the charges from a large number of conformers across the pseudorotational itinerary.

2.3 Methods

All the MD simulations were carried out using the AMBER 9.0 [109] suite of programs. We adopted the combined AMBER/GLYCAM force field for the simulations of **1**. For solution simulations, a 200 ns MD simulations of **1** were performed with the explicit

inclusion of a box of 298 TIP3P [110] water molecules under NPT conditions. The total box size was (25.569 × 25.372 × 25.544) (Å). The temperature was set to 300 K and the pressure to 1 atm. Non-bonded interactions were treated with a cutoff of 8 Å for solution simulation and 18 Å for gas phase simulations. The SCNB and SCEE scaling parameters were both set to unity in accordance with the GLYCAM approach. Prior to production MD simulations, minimization of the waters was first performed, followed by minimization of the entire system. The entire system was then annealed for 100 ps and equilibrated for 150 ps. Long-range electrostatic interactions were handled using Ewald summation. Bonds containing hydrogen were constrained to their equilibrium lengths using the SHAKE algorithm [24].

Two charge derivation procedures were considered to obtain atomic charges. The first one is the ensemble average approach proposed by Woods and workers [111] and is referred to as the usual GLYCAM procedure. Following this procedure, crystallographic data [112] were employed for the input geometry of methyl α-D-arabinofuranoside and an *ab initio* geometry optimization was then performed at the HF/6-31G* level of theory. All electronic structure calculations were performed using the Gaussian 03 software package [113]. Based on the HF/6-31G* single point, the RESP [51] approach was used to obtain an initial set of restrained partial atomic charges. A relatively short MD simulation (10 ns) based on these charges and one hundred conformations were selected from the resulting trajectory. The dihedral angles of the rotatable exocyclic moieties, such as hydroxyl groups, were then determined from the 100 snapshots and transferred to the quantum mechanics optimized geometry. Single point HF/6-31G* calculations were performed for these 100 new conformations. Partial atomic charges were obtained using the RESP approach for the 100 conformations and the final charge of each atom was obtained as an average. The value of the RESP restraint weight was set to 0.01 and the fitting was performed on all of the atoms except the aliphatic hydrogen [59]. The second charge derivation procedure is an important result of the current report and is described in Section 2.4.

2.4 Results

Atomic Charges. The atomic charges obtained from the standard GLYCAM procedure are shown in Table 2.1. The charges are calculated for five different ring conformers of **1**, labelled A-E. The structures of five reference rings are shown in Figures 2.5. It is clear from this data that the charges vary when one changes the ring conformation. While this variation is not large for all atoms, the effect is especially pronounced for atoms C3, C4 and C5. For example, for C3 the charges vary over the range 0.20-0.42. This variability will negatively impact the accuracy of the simulations. To remove the bias associated with the choice of a specific ring conformation, we developed a charge averaging procedure that accounts for the various furanoside ring conformations.

Ring-Averaged Charges. The usual GLYCAM approach is modified to obtain ring-averaged charges, which now incorporates the effects of the ring flexibility. Two hundred conformations were selected from a 50 ns simulation and a constrained *ab initio* geometry optimization (HF/6-31G*) was performed for each. During those constrained optimizations, the dihedral angles involving hydroxyl protons were held to the values obtained from the MD simulation. For each of the 200 new conformations, single point HF/6-31G* calculations were performed for the RESP fit. Note that the ring geometry and the dihedral angles involving hydroxyl protons are different in each of the 200 geometries. The same RESP approach as the one used in the usual GLYCAM procedure was then followed to obtain partial atomic charges. The charges obtained from our new procedure, where they are ensemble averaged over several exocyclic torsions *and* ring conformations, are presented in Table 2.1. We note that the new charges differ from those of the standard GLYCAM approach most notably for carbon atoms C3, C4, and C5. An average RMSD of the carbon atoms of the ring based on the 200 conformations used in the ring averaging was calculated and a value of 0.09 with a fluctuation of 0.08 was obtained. This parameter is a convenient measure of the ring flexibility of the system. Along with the calculation of the RMSD, a correlation study between RMSD and puckering was carried out

to quantify the magnitude of the RMSD in terms of puckering. In essence, this correlation study will indicate what change in ring puckering corresponds to a certain value of RMSD. However, this correlation study cannot be performed accurately on 200 conformations. It is necessary to consider many more conformations to get a statistically meaningful estimate. Therefore, we selected 100,000 conformations from the simulation based on our new ring averaged atomic charges, whose results will be shown and discussed below. Based on that study, the current average RMSD of 0.09 corresponds to a change of about 60 degrees in the puckering angle, P.

In the development of our ring average procedure, an alternate approach was attempted where one does not only freeze the dihedral angles involving hydroxyl protons (as in our final average ring procedure) but where one also freezes the dihedral angles of the ring (essentially fixing the ring puckering) in the geometry optimization of the 200 conformations selected from the simulation. In this way, the shape or puckering of the ring from the MD will be preserved and our ring average will be more consistent with the simulation, and therefore, with the flexibility of the system. However, the geometry optimization of the 200 conformations with all these constraints did not converge. The conformations were over constrained and all attempts to make them converge failed. The conformations extracted from the simulation seem to be very far from the *ab initio* minimum, and many constraints render convergence impossible.

Solution Simulations. Having determined the average atomic charges for **1**, we next set to establish the length of simulation required to achieve convergence. As a criteria for evaluating convergence we used the populations of rotamers about the C4-C5 bond. Shown in Figure 2.6 are the results of a convergence study of these rotamer populations in **1** as a function of simulation time. Charges obtained with the new ring-averaged procedure were used. From these results, it is clear that a 200 ns simulation is required to converge the populations of all the rotamers to reasonable uncertainties (a few units of percentage). Of particular note, simulations of less than 50 ns produced rotamer populations differing substantially from those obtained after 200 ns.

[c]

Table 2.1: Partial atomic charges of **1** obtained using the usual GLYCAM procedure for five reference rings (A-E) and using the averaged approach described here.

atom	A	B	C	D	E	ring averaged[a]
	P= 13[b]	P= 13	P= 32	P= 139	P= 58	P^*= 31
	ϕ_m= 34	ϕ_m= 41	ϕ_m= 40	ϕ_m=35	ϕ_m= 40	ϕ_m^*=35
C1	0.38 (0.05)[c]	0.37 (0.06)	0.38 (0.05)	0.37 (0.04)	0.38 (0.05)	0.38 (0.04)
C2	0.35 (0.09)	0.33 (0.09)	0.30 (0.07)	0.31 (0.05)	0.28 (0.09)	0.31 (0.07)
O2	-0.72 (0.02)	-0.73 (0.02)	-0.69 (0.02)	-0.70 (0.02)	-0.70 (0.03)	-0.69 (0.02)
OH2	0.42 (0.01)	0.43 (0.02)	0.42 (0.01)	0.43 (0.02)	0.42 (0.02)	0.42 (0.01)
C3	**0.34 (0.1)**	**0.42 (0.09)**	**0.24 (0.09)**	**0.20 (0.08)**	**0.39 (0.10)**	**0.30 (0.12)**
O3	-0.73 (0.03)	-0.76 (0.04)	-0.71 (0.03)	-0.73 (0.03)	-0.74 (0.02)	-0.72 (0.03)
OH3	0.43 (0.01)	0.43 (0.02)	0.43 (0.02)	0.44 (0.03)	0.43 (0.02)	0.43 (0.02)
C4	**0.19 (0.05)**	**0.12 (0.05)**	**0.33 (0.1)**	**0.40 (0.1)**	**0.18 (0.05)**	**0.26 (0.11)**
O4	-0.49 (0.04)	-0.47 (0.04)	-0.49 (0.05)	-0.46 (0.04)	-0.45 (0.04)	-0.47 (0.05)
C5	**0.32 (0.03)**	**0.31 (0.04)**	**0.22 (0.05)**	**0.20 (0.05)**	**0.28 (0.04)**	**0.24 (0.04)**
O5	-0.72 (0.03)	-0.67 (0.02)	-0.67 (0.02)	-0.69 (0.03)	-0.70 (0.02)	-0.67 (0.03)
OH5	0.42 (0.03)	0.41 (0.02)	0.42 (0.02)	0.43 (0.03)	0.42 (0.02)	0.42 (0.02)

[a]For the ring-averaged results, P^* and ϕ_m^* indicate the most probable values base on the distribution shown in Figure 2.8.

[b]Puckering angles, P, and amplitudes, ϕ_m, are calculated according to the Altona-Sundaralingam method [87].

[c]Numbers in parentheses correspond to standard deviations.

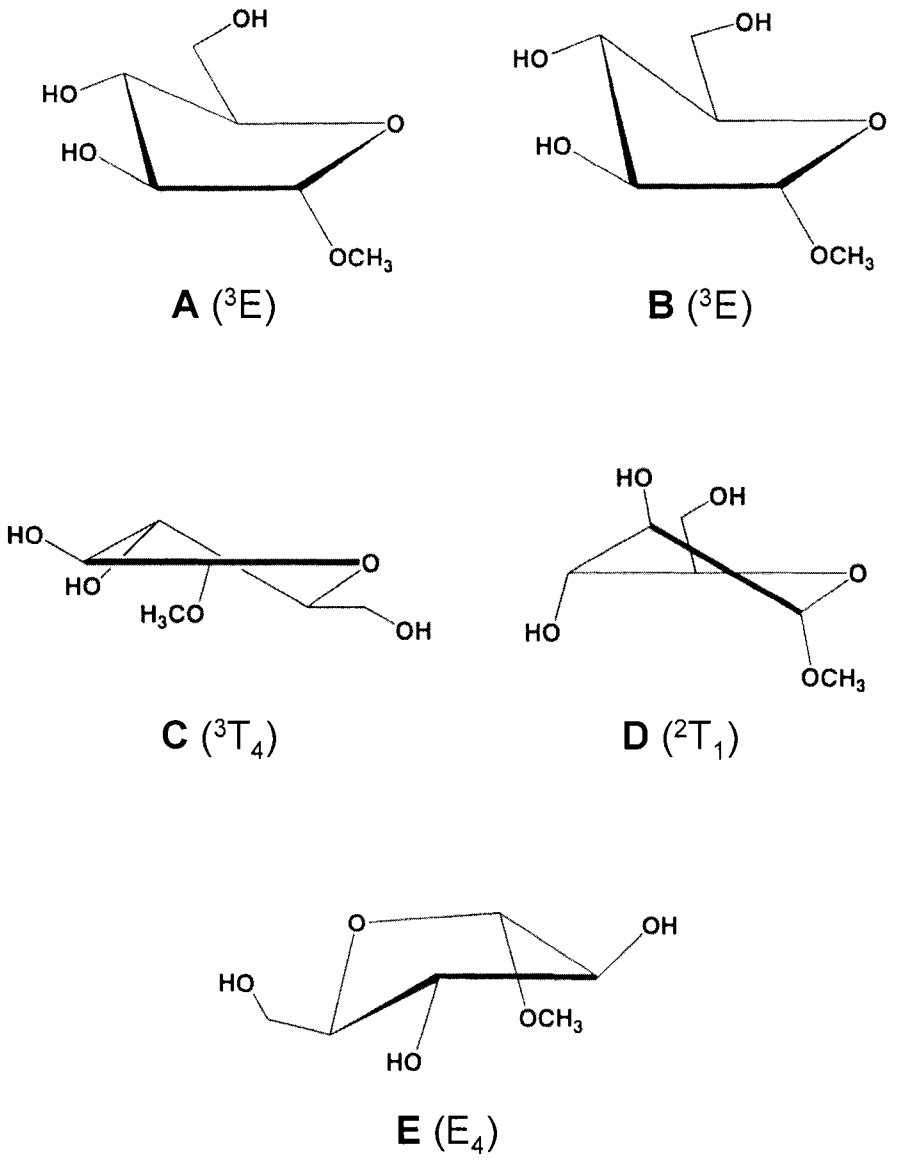


Figure 2.5: Structures of five reference rings (A-E).

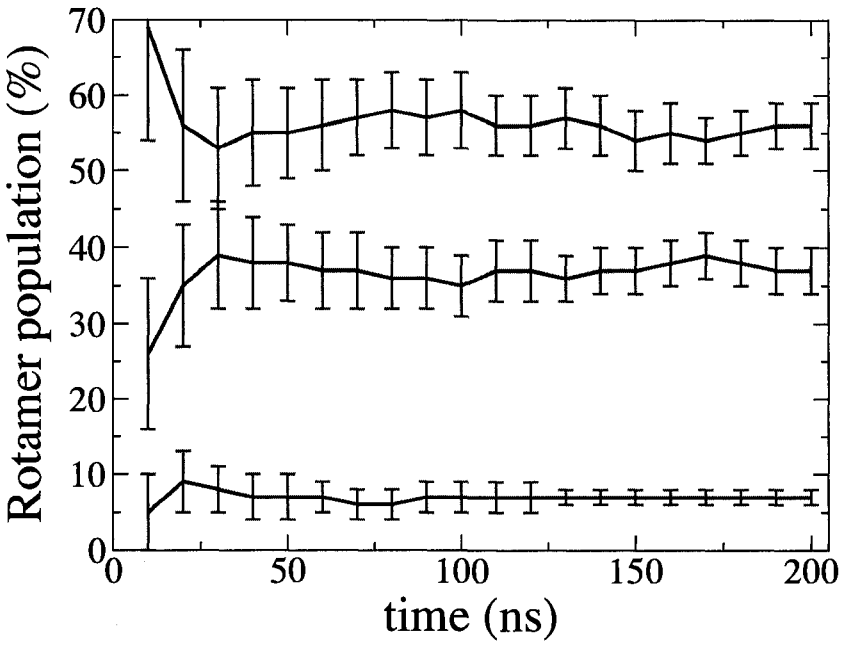


Figure 2.6: Convergence of the rotamer populations of **1**. Lines are a guide to the eye, and the *gg*, *gt* and *tg* populations are given by the top, middle, and bottom lines, respectively.

We next compared the C4-C5 rotamer populations obtained from the simulations with those derived from experimental results [96]. A histogram of the behaviour of this torsion is shown in Figure 2.7. All three rotamers are populated but the *tg* rotamer (180°) is visited infrequently. When the conformers from the three peaks in the histogram are integrated, it is possible to quantify rotamer populations, which are presented in Table 2.2. In addition to the results based on our ring-averaged charge derivation procedure and the experimental values, the results of simulations based on the five charge sets of the standard (fixed ring) GLYCAM procedure are also presented. Clearly, the new ring-averaged charge calculation procedure leads to a good agreement with experiment, which is better than the fixed ring method. While both charge derivation approaches yield the correct ordering of the rotamer populations, the results based on the usual GLYCAM approach can sometime lead to

Table 2.2: Rotamer populations of **1** obtained using the various approaches.

rotamer population (%)	*gt*	*tg*	*gg*
experiment [96]	38	14	48
ring average charges	37(3)	7(1)	56(3)
fixed ring charges A	29(2)	8(1)	63(3)
fixed ring charges B	29(2)	8(1)	63(3)
fixed ring charges C	39(3)	7(1)	54(3)
fixed ring charges D	33(3)	8(1)	59(3)
fixed ring charges E	27(2)	8(2)	65(3)
gas phase	7(1)	40(3)	53(3)

a worse agreement with experiment because of the intrinsic ring bias of that procedure. These results validate the ring-averaging method for obtaining charges in these flexible rings and, encouraged by these results, we considered other ring parameters in **1**, in particular P and ϕ_m.

Figure 2.8 contains the variation in P, which describes ring puckering; the inset shows the variation in puckering amplitude, ϕ_m. The distribution in ϕ_m is centered about 35°, which corresponds well to earlier *ab initio*, density functional theory and molecular calculations [98, 97, 99, 100, 102] on **1**, as well as to the puckering amplitude of this molecule in the crystal structure [112]. With regard to P, conformations with values in the northern hemisphere of the pseudorotational itinerary (see Figure 2.1) are clearly favoured although a small fraction of the conformers is also present in the southern hemisphere. The area of conformational space centered about $P = 45°$ corresponds well to the N conformer determined for **1** [96] using the PSEUROT [90] procedure, which identified two conformers: a N conformer at $P = 44°$ (39 %) and a S conformer at $P = 123°$ (61 %). However, while there is good agreement with the identification of the N conformer, the conformer populations obtained from the simulation do not correspond well with experiment, nor with previous *ab initio* and density functional theory calculations on **1** [98, 97, 99, 100, 102]. Indeed, the distribution shown in Figure 2.9 suggests that a while a small population of S conformer (centered around $P = 180°$) is present, the equilibrium is heavily biased to the N con-

former. These results suggest that the two-state model inherent in the PSEUROT approach may not be valid for **1**.

Figure 2.9 illustrates the correlation study mentioned earlier where we calculate the joint probability distribution of the puckering angle, P, and the RMSD of the ring atoms. The graph shows that a change of 180° in ring puckering, which is the maximum possible, represents a variation of approximately 0.25 in RMSD. The figure also reveals that an RMSD value of 0.09 as obtained in the ring averaged charge derivation procedure of the preceding section corresponds to a 60° change in the puckering angle, P. If the fluctuation magnitude of 0.08 is taken into account, the change in ring puckering will be more than 100°. Obviously, this result lends weight to our modification to the standard GLYCAM procedure to derive the set of atomic charges. The current solvated molecular system is very flexible and the charge derivation cannot be based on only one ring, it has to be based on an average over numerous rings to represent all the conformations accessible to the system.

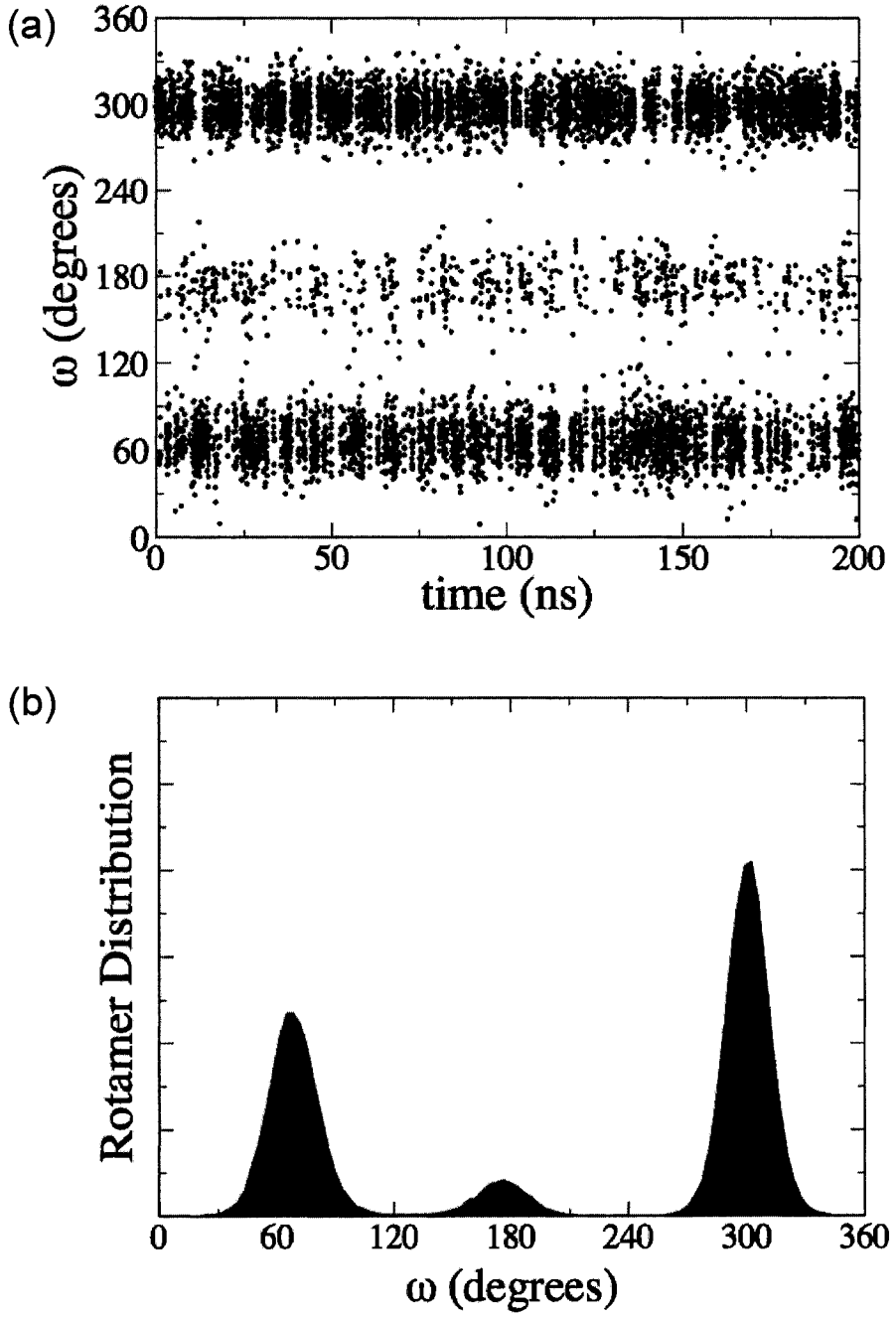


Figure 2.7: Time dependence of (a) the C4-C5 torsion angle and (b) its associated distribution for **1** in solution.

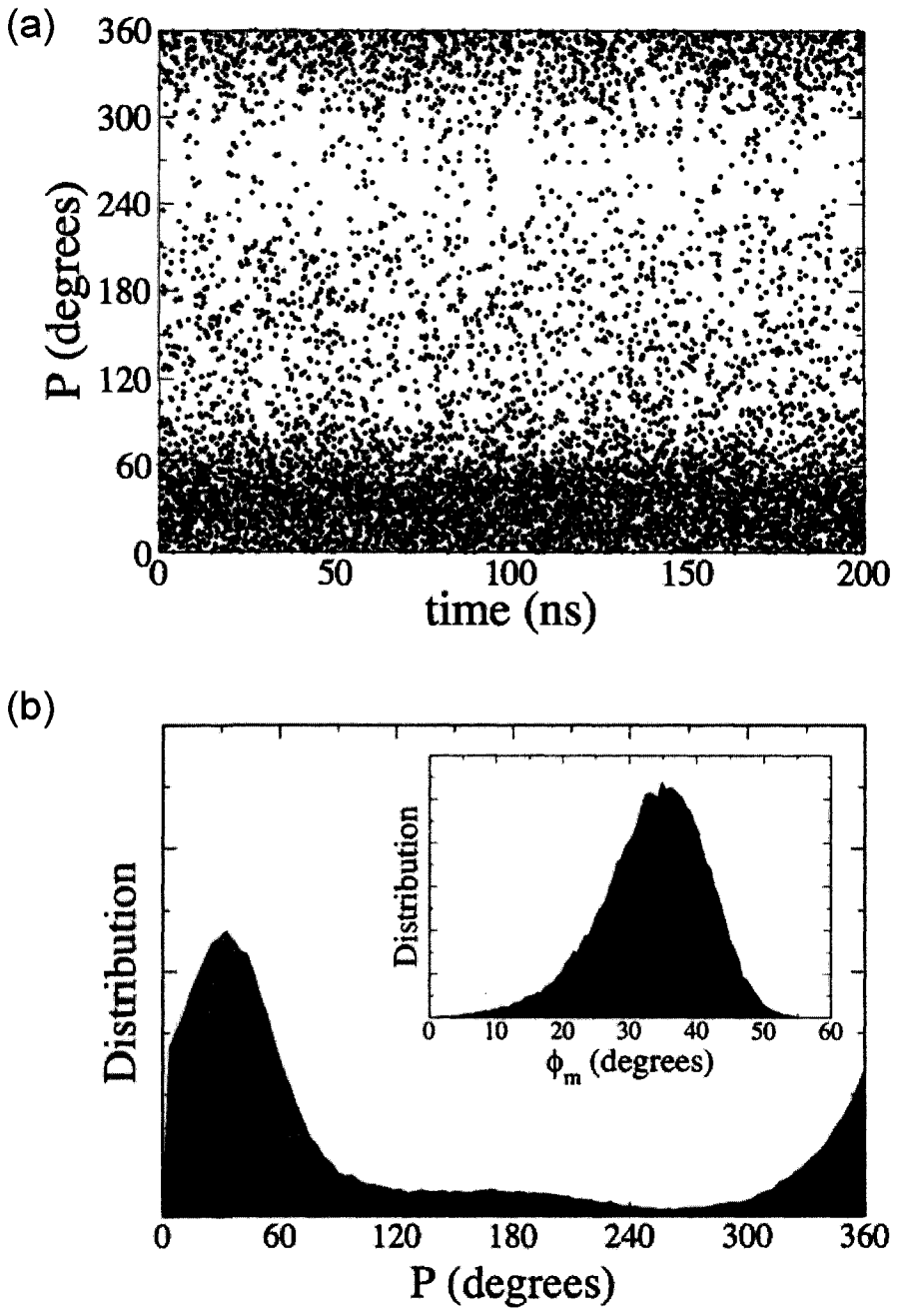


Figure 2.8: Time dependence of (a) the Altona-Sundaralingam P angle and (b) its associated distribution for **1** in solution. The distribution of puckeing amplitude, ϕ_m is given in the inset of the bottom panel (ϕ_m^*).

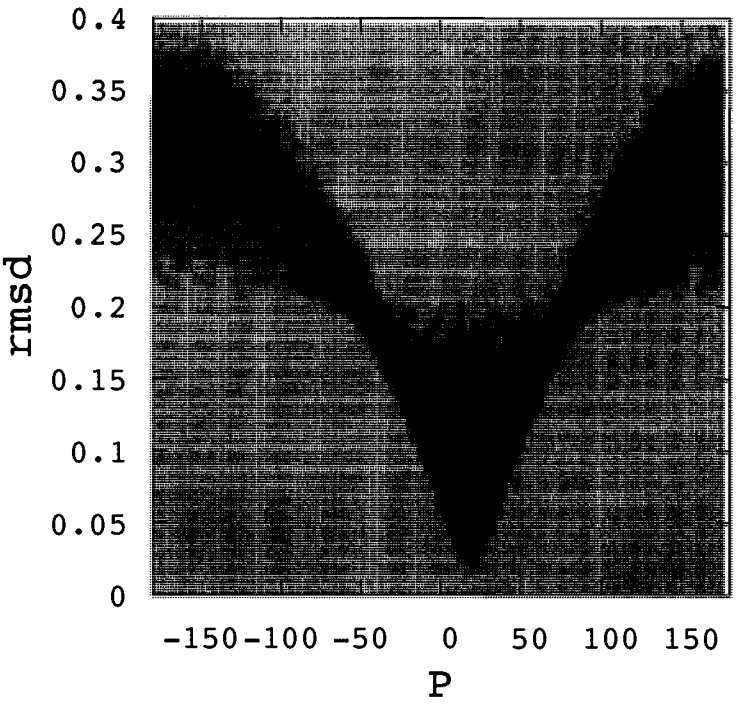


Figure 2.9: Joint probability distribution of the puckering angle, P (in deg), and the rmsd (Å) of the ring carbon atoms.

Gas-Phase Simulations. Although we anticipated that the inclusion of explicit water molecules to simulate solvent effects would be essential to obtain results consistent with experiment, as a test of this we performed a simulation of **1** in the gas phase. We present in Figure 2.10 and in Figure 2.11 the analysis of the C4-C5 torsion angle and, pseudorotation behaviour in the gas phase, respectively. As expected, these gas phase results differ from those obtained with explicit solvent inclusion. This is presumably due, in large part, to the fact that in the absence of water, the possibility of intermolecular hydrogen bond competition with the solvent is no longer possible.

We see from Figure 2.10 that the ordering of the rotamer populations is reversed compared to the solution and experimental cases. The population of the *tg* rotamer is now greatly enhanced at the expense of the *gt* rotamer. Figure 2.11 in turn reveals that the pseudorotation distribution now shows more distinct north (N) and south (S) populations. The most populated values of the two puckering states are $P_N^* = 38°$ and $P_S^* = 165°$ for the north and south regions, respectively, which agrees well with previous *ab initio* and density functional theory calculations on **1** [98, 97, 99, 100, 102]. This result differs significantly from the simulation done in the presence of water, where two distinct puckering states did not exist and instead a single region in the northern hemisphere of the pseudorotational itinerary was favoured. As expected, these results underscore the importance of using an explicit solvent model to correctly describe solution behaviour. An *ab initio* and density functional theory study of several conformers of **1** in the gas phase [97], showed high correlation between the rotamer and the ring puckering distributions. In other words, the rotamer population depends on the ring puckering and vice versa.

Motivated by this study, we carried out a correlation study between the C4-C5 torsion and the puckering angle. Figure 2.12 shows the joint probability distribution of the C4-C5 torsion and puckering angle, P, for both gas and solution phase simulations. The gas phase results reveal the presence of north and south hemispheres of the pseudorotational wheel, and different trends of C4-C5 torsion distribution are obtained for each hemisphere. For example, conformations with P values between 0° and 50° (North) exhibit the trend in rotamers of $tg > gg > gt$, whereas for confor-

mations with P values around 180 degrees (South), the trend is $gg > tg = gt$. The favouring of the gg rotamer in the S conformers would be expected given the ability of conformers with this C4-C5 torsion to form trans-annular hydrogen bonds between OH2 and OH5. Similarly, the tg rotamer is stabilized by hydrogen bonding between OH3 and OH5 in the N conformers. Therefore, there is a marked correlation between C4-C5 torsion and ring puckering in the gas phase, as concluded from an earlier *ab initio* study [97] although the trends in rotamers for the respective values of ring puckering do not coincide. The *ab initio* study shows $gg > gt > tg$ for $P \approx 30°$ and $gg > tg > gt$ for $P \approx 180°$. These differences may arise from the fact that in the *ab initio* study a full sampling of conformational space was not undertaken. Instead the energy-minimized structures were obtained by full optimization of a family of 30 ring-constrained conformers [98] that had been partially optimized to probe the effect of ring conformation on various molecular parameters, *e.g.*, bond-lengths and bond angles. In solution, this strong correlation between C4-C5 rotamer and furanose ring conformation is not observed. As seen in Figure 2.12, the north hemisphere of the pseudorotational wheel is mostly populated, regardless of C4-C5 rotamer. We propose that the effect is due to the lack of intramolecular hydrogen bonding in the solution simulations.

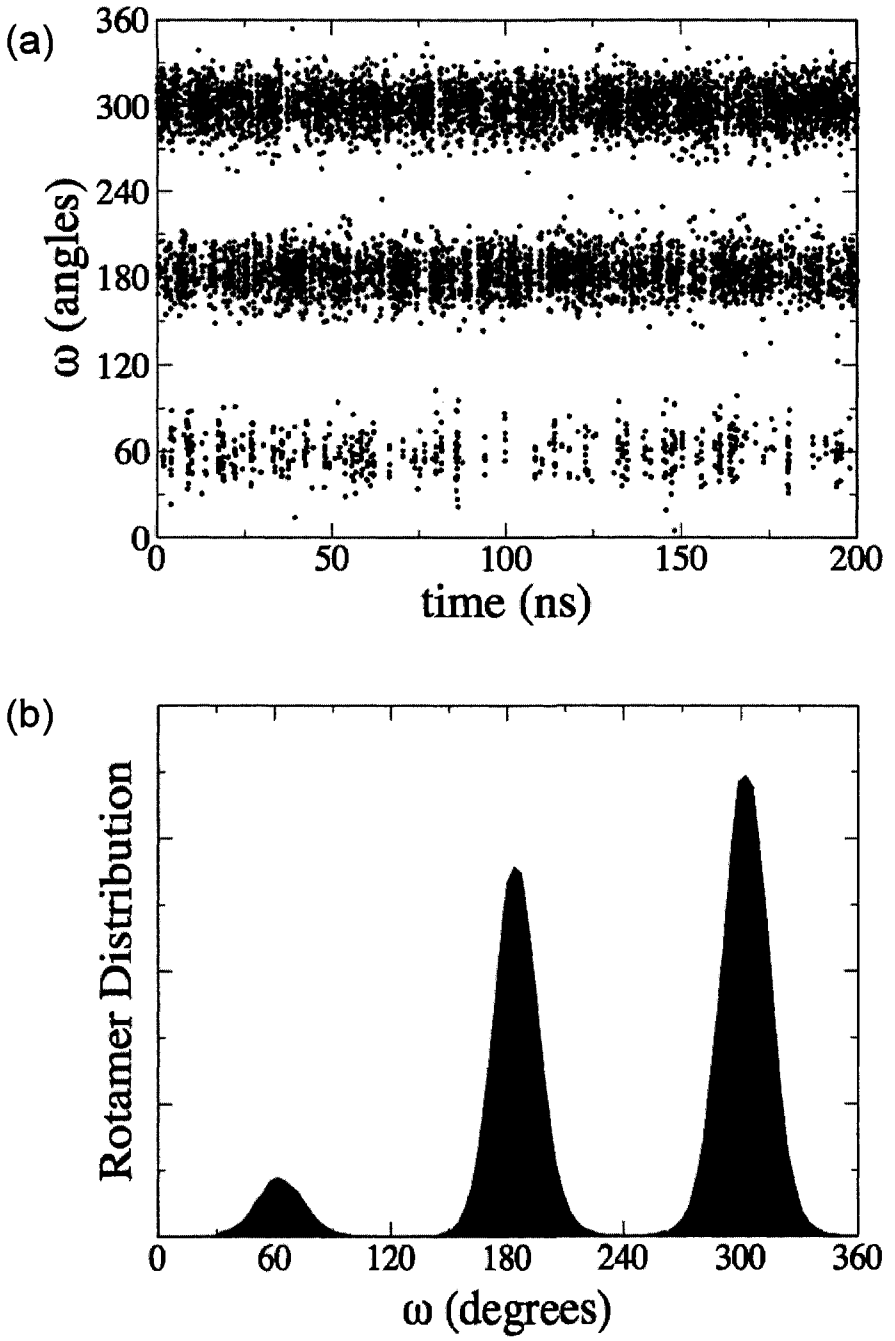


Figure 2.10: Time dependence of (a) the C4-C5 torsion angle and (b) its associated distribution for **1** in the gas phase.

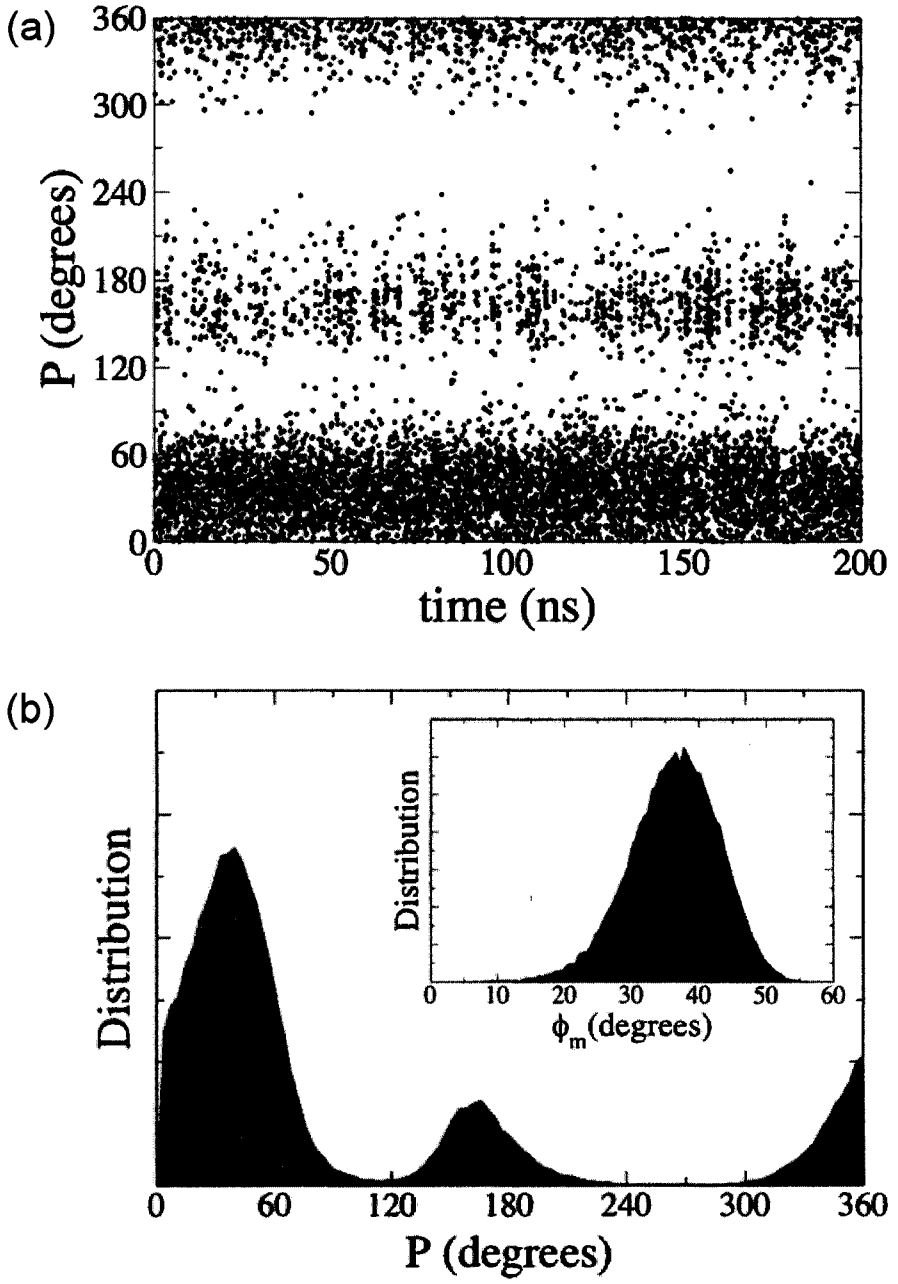


Figure 2.11: Time dependence of (a) the P angle and (b) its associated distribution for **1** in the gas phase ($P_N^* = 38$ and $P_S^* = 165$). The distribution of puckeing amplitude, ϕ_m is given in the inset of the bottom panel ($\phi_m^* = 38$).

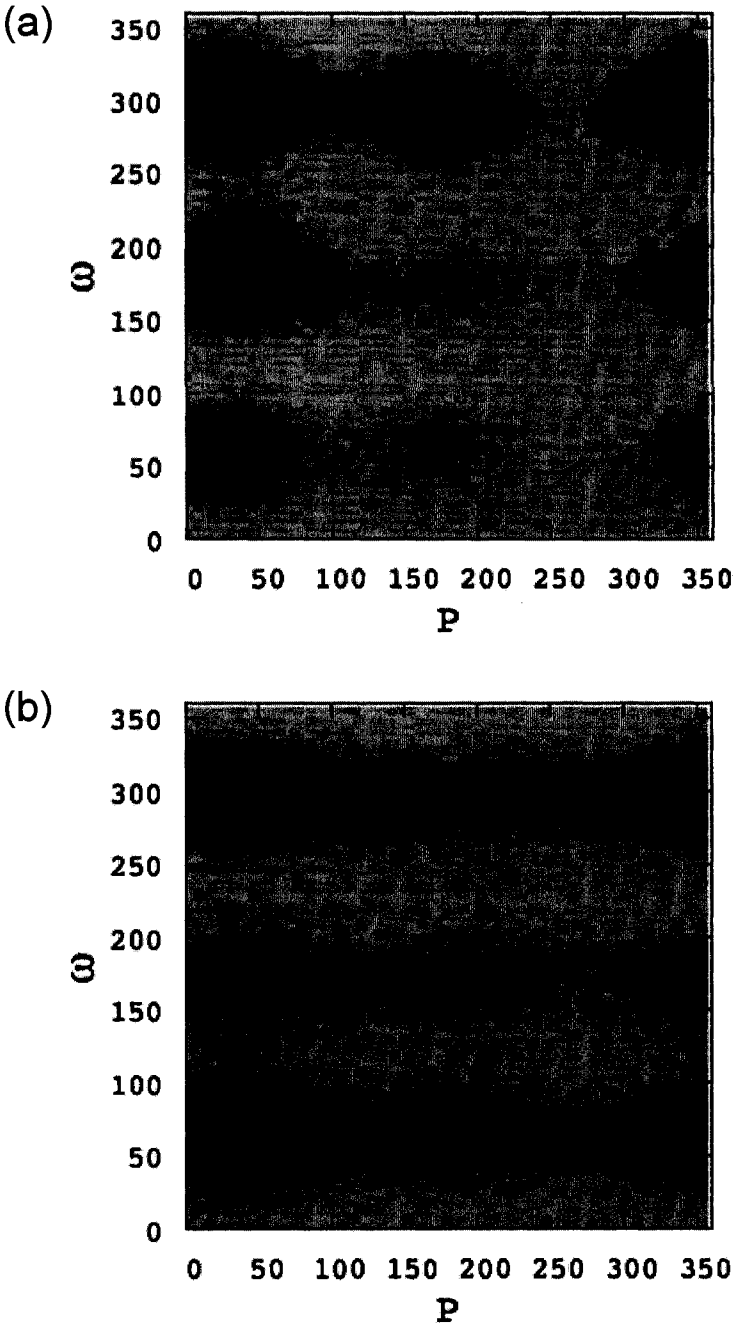


Figure 2.12: Joint probability distribution of the puckering angle, P, and the C4-C5 torsion for **1** in (a) the gas and (b) solution phases. The units of the angles P and ω are in deg.

2.5 Conclusion

In this study, a new AMBER/GLYCAM approach for deriving atomic charges has been suggested, and applied to methyl α-D-arabinofuranoside (**1**). The major difference that distinguishes this new method from the previously suggested method is the consideration of the inherent flexibility of five-membered rings. The usual GLYCAM procedure that has been thought to be a good standard for deriving charges and therefore has been frequently used, assigns atomic charges based on only a single ring conformer. The standard procedure is shown to be applicable to the models for rigid rings (e.g. pyranosides), but not to the ones for flexible rings such as furanosides. However, this study takes the floppiness of furanosides into account and derives charges by averaging them over a large number (two hundred) of selected conformers in an MD simulation.

Furthermore, long simulation times (200 ns) are required to achieve convergence. Rotamer population about C4-C5 bond and puckering amplitude of the ring (ϕ_m) obtained from this computational study agree well with the experimental results from NMR spectroscopy. Simulation results from this study on a furanoside ring in aqueous environment revealed only one low energy region in conformational space rather than two. Hence, the popular two-state model for the conformation of furanoside rings may not be valid in the case of **1** in water, whereas the simulation results in gas phase as well as the previous studies of *ab initio* and DFT [98, 97, 99, 100, 102] all support the validity of the two-state model. This suggests that one must take extreme care when applying the two-state model since the two-state model may be unreliable in some situations, and that the valid limits of the model must be further explored. This work provides a stepping stone to studies to come in the future.

Chapter 3

Intermolecular Interactions within Desolvated Protein-Ligand Complexes

3.1 Introduction

Most biological processes, including the immune response, bacterial and viral infections, involve the association of biomolecules to form specific, noncovalent complexes. The structure and stability of these complexes are determined by the concerted action of many forces (e.g. hydrogen bonds (H-bonds), ionic and van der Waals interactions) between binding partners and from the displacement and reorganization of solvent molecules associated with the solvent shell of the binding partners. An understanding of these forces, the thermochemistry and the structures they lead to, is essential to a complete understanding of biological processes. In addition, since biochemical function is typically mediated by these forces present in biomolecules, an understanding the links between structures and functions of biological complexes provides an

approach to understanding the origin of disease and the effects of therapy at the molecular and cellular level.

To further develop this understanding, a multi-disciplinary approach that combines biology, chemistry, biophysical chemistry and physics is required due to the structural complexity of biomolecules. A variety of different methodologies and techniques such as NMR spectroscopy, isothermal titration calorimetry and X-ray crystallography are used to characterize noncovalent complexes [114, 115, 116].

Gas phase studies of desolvated biological complexes represent a promising experimental approach to probe directly the intrinsic (solute-solute) intermolecular interactions and, indirectly, the role of solvent in biological recognition. The transfer of specific, noncovalent biological complexes from solution to the gas phase is, in most cases, readily achieved using electrospray ionization (ES) [117]. Because the interactions between biological macromolecules and individual water molecules are typically weak, the hydration waters are rapidly lost in the gas phase giving the desolvated ions. Once in the gas phase, the ions can be interrogated using a variety of mass spectrometry (MS)-based techniques. Additional experimental techniques used for this study will be described in Section 3.2.

Elucidating the higher order structures of gaseous ions of large biological molecules and their noncovalent complexes represents a significant experimental challenge. Many studies have shown the evaluation of higher order structures of gaseous biopolymers from various experimental techniques [118, 119, 120, 121, 122, 123]. Individual intermolecular interactions within gaseous ions of noncovalent biological complexes can also be inferred. Normally, they are determined from differences in the stability, usually kinetic, of structurally-related complexes. There are several examples to identify the ligand binding site [124] and evaluate the binding interactions [125] within the gaseous ions of protein-ligand complexes.

Recently, Kitova and co-workers developed a reactivity-based approach, employing blackbody infrared radiative dissociation (BIRD) [126, 127], a thermal dissociation technique implemented with a Fourier-transform ion cyclotron resonance mass spectrometer (FT-ICR MS), and functional group replacement (FGR). An attractive

feature of the BIRD/FGR method is that it allows intermolecular interactions to be identified and quantified [128, 129]. However, the BIRD/FGR method has a limitation that it requires sufficient affinity in solution to lead to detectable concentrations of complex for the structurally-modified proteins and ligands through FGR. The disruption of certain key intermolecular interactions in the complex leads to a complete loss of binding in solution and, consequently, these interactions can not be investigated using the BIRD/FGR method. Again, the further description of this method will be provided in Section 3.2.

In present work, we seek to provide a complete description of the intermolecular interactions (H-bonds) within a desolvated noncovalent protein-ligand complex using both experimental and computational methods. MD simulations have been used extensively to evaluate the higher order structure of gaseous ions of peptides, proteins, as well as other biological molecules and their noncovalent complexes [130, 131, 132, 133, 134]. Here, MD simulations were performed on the gaseous protein-ligand ions to complement the gas phase measurements. MD simulations play a dual role in this investigation: they provide a means of confirming interactions identified from experimental techniques as well as predicting additional sites of interaction on the complex, including sites for which the binding partners cannot be experimentally determined. Taken together, the results of this study provide the first detailed and quantitative description of the intermolecular interactions within the gaseous ions of a protein-ligand complex.

In the following section, fundamental aspects of the experimental techniques used in this study and the experimental results will be briefly described.

3.2 Experimental Methods and Summary of Experimental Results

Mass spectrometry (MS) has been widely used in biomedical research as it offers advantage in sensitivity, speed, specificity and accuracy of mass determination [135]. MS combined with electrospray (ES) [136, 137] or nanoelectrospray (nanoES) [138, 139]

ionization has emerged as a powerful tool for studying the complexation processes of noncovalent biomolecular complexes such as protein assemblies, protein-ligand complexes, oligonucleotide duplexes in solution. Various biochemical information is available from ES-MS experiments: detection of specific biomolecular complexes in solution, direct determination of their binding stoichiometry [117, 140, 141] and measurements of relative [142, 143, 144] and, in some cases, absolute [145, 146] binding affinities.

Beyond its ability to transfer biomolecular complexes in solution to the gas phase in an ionized form, ES-MS, in conjunction with gas-phase dissociation techniques, ES-MS becomes possible for identifying binding sites and investigating intrinsic noncovanlent interactions [147, 148, 129, 128]. Many studies have suggested that at least some aspects of the higher order structure of proteins [149] or to some extent, specific intermolecular interactions in protein-ligand complexes can be preserved after transfer into the gas phase [128, 147]. Thus, the ES-MS observations of gaseous noncovalent complexes reflect the nature of the interactions found in solution [150].

Elucidation of the intermolecular interactions present in gaseous protein-ligand complexes can be realized by employing blackbody infrared radiative dissociation (BIRD) [126, 127], a thermal dissociation method combined with a Fourier-transform ion cyclotron resonance mass spectrometer (FT-ICR MS), and functional group replacement (FGR) [128, 129]. Using this approach, individual interactions in gaseous biological complexes can be identified and the strength of the interactions can be quantified. In the BIRD/FGR method, to identify whether a particular functional group, either on the protein or ligand, is involved in binding, the group is modified in such a way that any pre-existing interaction is lost. The activation energy, E_a, of the modified complex is then compared to the E_a of the unmodified complex. The value of E_a is determined from the slope of the linear least-squares fit of Arrhenius plot, which constructed from the temperature-dependent rate constants measured from the BIRD method. A decrease in E_a upon modification indicates that the particular functional group stabilized the complex. Furthermore, the difference in E_a, *i.e.* $\Delta E_a = E_a$ (unmodified complex) - E_a (modified complex), provides a measure of the

strength of the interaction. To identify H-bond donor/acceptor pairs, a three step approach is normally utilized, in which the ΔE_a values are determined for complexes containing a single modification of the ligand (functional group modification), a single modification of the protein (active site mutation) and simultaneous modification of the protein and the ligand (dual modification) [128, 129]. For a given donor/acceptor pair, the magnitude of the ΔE_a values determined for all three complexes will be identical.

Several preliminary studies [128, 129, 151] described the application of the BIRD/FGR technique to a gaseous protein-trisaccharide complex consisting of a genetically engineered single chain variable fragment, scFv, of the monoclonal antibody Se155-4 and its native trisaccharide ligand, αGal[αAbe]αMan (**1**). A representation of the 3D structure of the complex is shown in Figure 3.1. Crystal structures for the (scFv + **1**) complex [13] and the corresponding antigen binding fragment (Fab) complex have been solved. Analysis of the crystal structures suggests that **1** is bound

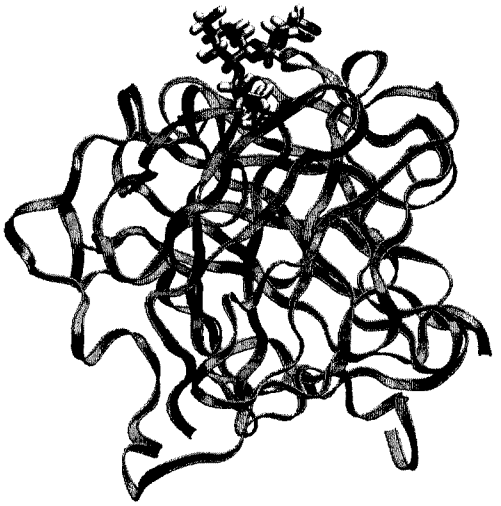


Figure 3.1: Structure of the complex of scFv and its trisaccharide ligand (**1**)

to the scFv through as many as five intermolecular H-bonds in solution (see Figure 3.2). Additionally, a water molecule (Wat1) at the base of the binding site, which mediates H-bonds between scFv and **1**, has been identified. Two additional waters (Wat2, Wat3) are also observed in the crystal structures. It has previously been

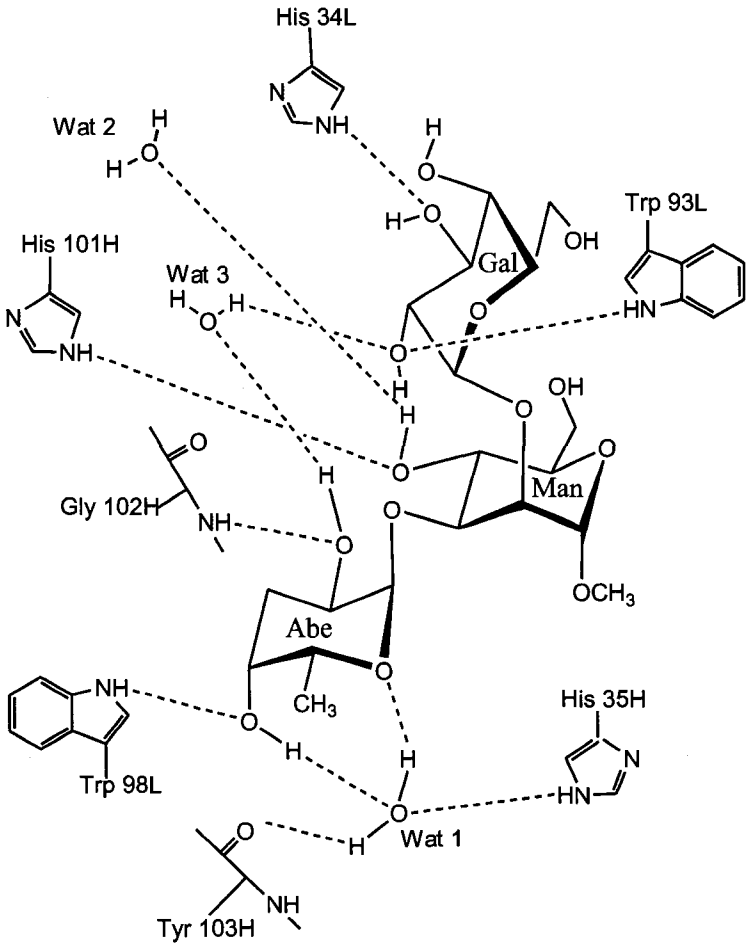


Figure 3.2: Intermolecular hydrogen bond scheme for the complex of scFv and its trisaccharide ligand (**1**) obtained from X-ray analysis of the crystal structure.

shown that one of the specific intermolecular H-bonds (His^{101H} - Man C4 OH) is preserved in the gas phase, at least at certain charge states [129, 151]. Indirect evidence for the formation of nonspecific interactions, *i.e.* interactions not present in solution but which form in the gas phase, was also reported [128].

In the present work, the BIRD/FGR technique was applied to the complexes of **1** and monodeoxy analogs (**2** - **5**) with the scFv and an array of single point scFv mutants to provide a complete description of the intermolecular interactions within the protonated and deprotonated ions of the desolvated (scFv + **1**) complex.

The structures of trisaccharide ligands are shown in Figure 3.3: αGal[αAbe]αMan (**1**), (3-deoxyαGal)[αAbe]αMan (**2**), (6-deoxyαGal)[αAbe]αMan (**3**), αGal[αAbe](4-deoxyαMan) (**4**), and αGal[αAbe](6-deoxyαMan) (**5**). A series of single point scFv mutants were prepared by using site-directed mutagenesis: $His^{101H}Ala$, $His^{101H}Arg$, $His^{101H}Lys$, $His^{101H}Gln$, $His^{34L}Ala$, $His^{35H}Ala$, $His^{97L}Ala$, $Trp^{33H}Ala$, $Trp^{33L}Ala$, $Trp^{98L}Ala$, $Asp^{96L}Ala$. Measurements were performed over a range of charge states in order to asses the influence of charge on the nature and strength of the intermolecular interactions.

To provide a complete description of the intermolecular interactions within the protonated and deprotonated ions of the desolvated (scFv + **1**) complex, MD simulations were performed and the intermolecular H-bonds were identified. Since MD simulations were performed on the $(\text{scFv} + \mathbf{1})^{n+/-}$ ions at charge state +8 and -8, we only report here the experimental data obtained from BIRD/FGR technique for the $(\text{scFv} + \mathbf{1})^{n+/-}$ ions at charge state +8 and -8. Arrhenius parameters (activation energies and A-factors) determined for the dissociation of the $(\text{scFv} + \mathbf{1})^{8+/-}$ ions are listed in Table 3.1 and 3.2. Figure 3.4 shows the summary of interaction maps determined from BIRD/FGR method for the $(\text{scFv} + \mathbf{1})^{8+/-}$ ions. These maps are compared to the maps (see Figure 3.6) obtained from MD simulations in the Section 3.4. Also, comparison of the interactions identified in the gas phase with the H-bond map inferred from crystallographic data provides new insights into the structural changes that accompany the transfer of protein-ligand complexes from solution to the gas phase by ES and the influence of charge state thereon.

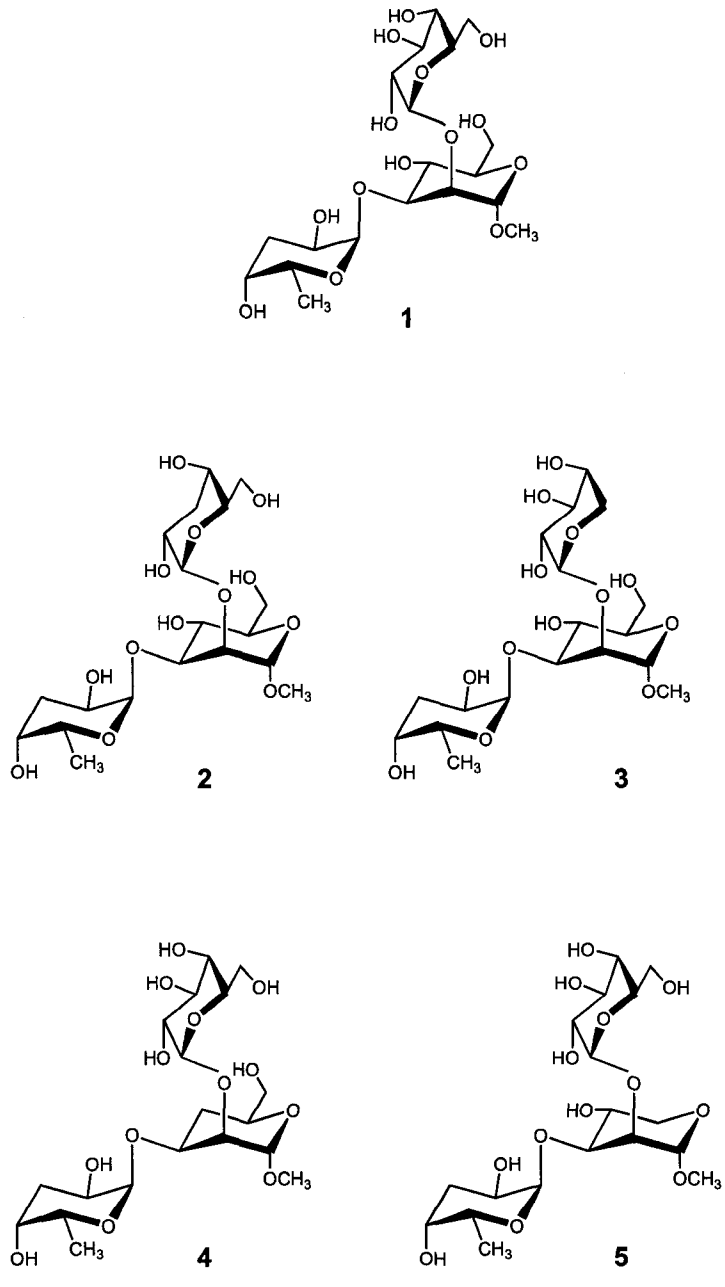


Figure 3.3: Structures of the trisaccharide ligands (**1** - **5**).

Table 3.1: Arrhenius Parameters determined for the dissociation reaction: (scFv + **1**)$^{8+/-}$ $\rightarrow$ scFv$^{8+/-}$ + L, where L = **1** - **5**.

ligand	charge state	E_a(kcal/mol)	$\Delta_L E_a$(kcal/mol)	A (s^{-1})
1	-8	49.5 $\pm$ 0.5[a]	-	$10^{25.0\pm0.3}$
	+8	54.9 $\pm$ 1.5	-	$10^{27.6\pm0.8}$
2	-8	49.0 $\pm$ 1.1	0.5 $\pm$ 1.2	$10^{25.4\pm0.6}$
	+8	48.5 $\pm$ 0.9	6.4 $\pm$ 1.7	$10^{25.0\pm0.5}$
3	-8	48.1 $\pm$ 0.7	1.4 $\pm$ 0.9	$10^{25.0\pm0.4}$
	+8	47.6 $\pm$ 0.2	7.3 $\pm$ 1.5	$10^{24.2\pm0.1}$
4	-8	45.0 $\pm$ 0.5	4.5 $\pm$ 0.7	$10^{23.2\pm0.3}$
	+8	51.7 $\pm$ 0.9	3.2 $\pm$ 1.7	$10^{26.4\pm0.5}$
5	-8	49.0 $\pm$ 1.3	0.5 $\pm$ 1.4	$10^{25.5\pm0.7}$
	+8	46.9 $\pm$ 1.5	8.0 $\pm$ 2.1	$10^{24.4\pm0.8}$

[a]Errors are on standard deviation.

Table 3.2: Arrhenius Parameters determined for the dissociation of (scFv + **1**)$^{8+/-}$ ions composed of the trisaccharide ligands, L = **1** - **5** and scFv mutants.

mutant	ligand	charge state	E_a(kcal/mol)	$\Delta_P E_a$(kcal/mol)	$\Delta_{PL} E_a$(kcal/mol)	A (s^{-1})
$His^{101H}Ala$	**1**	-8	47.1 ± 0.8[a]	2.4 ± 0.9	-	$10^{24.0\pm0.4}$
	1	+8	52.1 ± 1.0	2.8 ± 1.8	-	$10^{26.2\pm0.5}$
	4	-8	45.2 ± 10.7	-	4.3 ± 0.9	$10^{23.4\pm0.4}$
	4	+8	52.0 ± 0.6	-	2.9 ± 1.6	$10^{26.6\pm0.3}$
$His^{34L}Ala$	**1**	-8	49.7 ± 1.0	-0.2 ± 1.1	-	$10^{24.9\pm0.5}$
	1	+8	50.9 ± 0.4	4.0 ± 1.6	-	$10^{25.5\pm0.2}$
	2	-8	45.6 ± 1.0	-	3.9 ± 1.1	$10^{23.4\pm0.5}$
	2	+8	47.2 ± 1.0	-	7.7 ± 1.8	$10^{24.3\pm0.5}$
$Asn^{96L}Ala$	**1**	-8	51.2 ± 0.8	-1.7 ± 0.9	-	$10^{25.6\pm0.4}$
	1	+8	52.6 ± 0.6	2.3 ± 1.6	-	$10^{26.4\pm0.4}$
	3	-8	48.1 ± 0.8	-	1.4 ± 0.9	$10^{25.0\pm0.4}$
	3	+8	49.2 ± 0.6	-	5.7 ± 1.6	$10^{24.2\pm0.3}$
$His^{35L}Ala$	**1**	-8	48.3 ± 0.5	1.2 ± 0.9	-	$10^{24.6\pm0.3}$
	1	+8	47.5 ± 1.1	7.4 ± 1.9	-	$10^{23.9\pm0.6}$
$His^{97L}Ala$	**1**	-8	49.4 ± 0.3	0.1 ± 0.6	-	$10^{24.9\pm0.2}$
	1	+8	50.7 ± 1.0	4.2 ± 1.8	-	$10^{25.5\pm0.5}$

[a]Errors are one standard deviation.

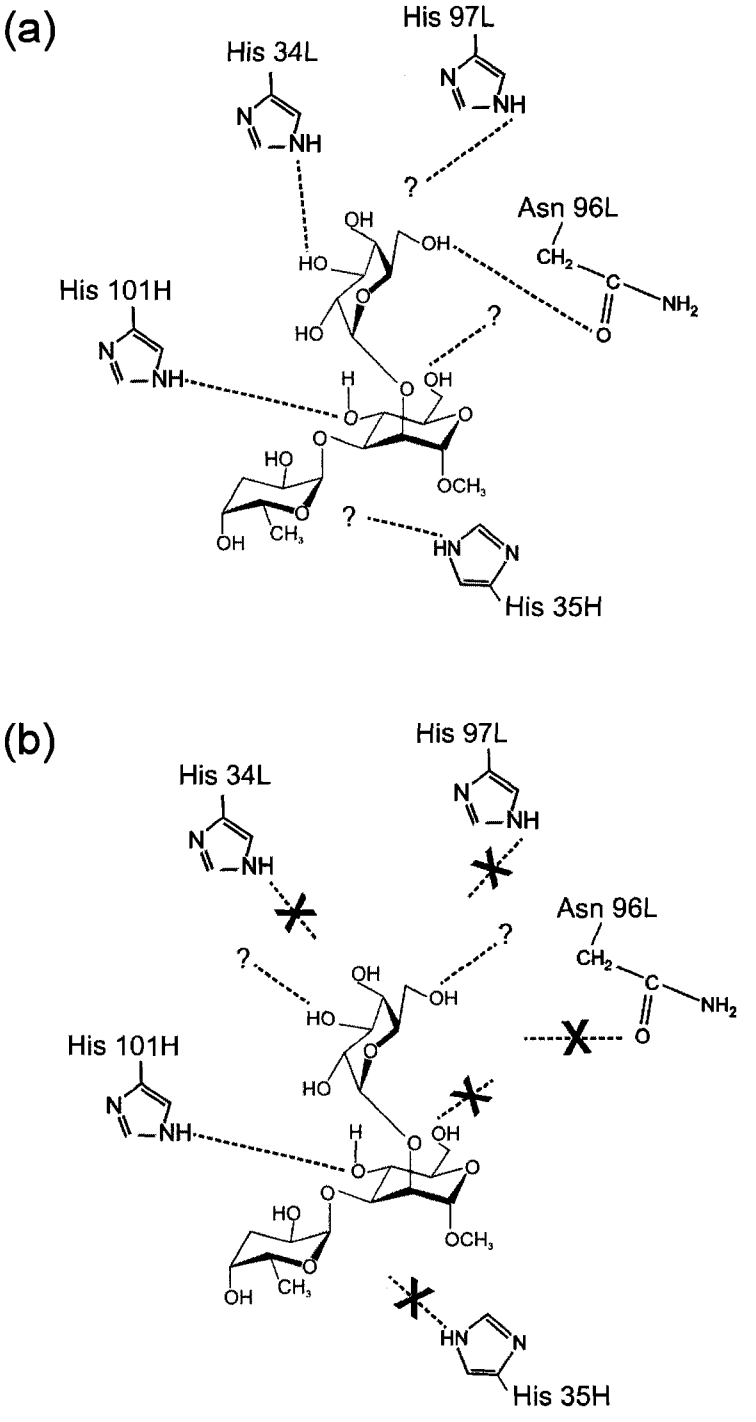


Figure 3.4: Interaction maps determined from BIRD/FGR data for the (scFV + $\mathbf{1})^{n+/-}$ ions at charge state (a) +8 and (b) -8.

3.3 Computational Methods

The AMBER 9.0 program suite[109] was used in MD simulations. The crystal structure (1MFA) [13] was used for the initial geometry of the (scFv + **1**) complex. The simulations were performed using the AMBER 94 forcefield with the GLYCAM parameter set for oligosaccharides [12]. Electrostatic potential (ESP) atomic partial charges, determined by Woods and coworkers [60], were used for **1**. The (scFv + **1**) complex at the +8 and -8 charge states were chosen for investigation. As described in more detail below, ten different charge distributions were considered for each charge state. The charge distributions considered for -8 charge state required that Arg residues be in the neutral form. Currently with AMBER, atomic charges and (atom type) parameters are only available for the Arg residue in its protonated form. Consequently, it was necessary to develop charges and parameters for the neutral form of Arg. The energies of the desolvated (scFv + **1**)$^{8+/-}$ ions were first minimized with the conjugate gradient method using a 0.0001 kcal/mol convergence criterion. The entire system was then heated from 10 to 300 K over a period of 15 ps. In order to mimic experimental conditions, simulations were performed in the gas phase under isothermal conditions. Constant temperature was maintained using the weak-coupling algorithm with time constant 1.0 ps [152]. During the simulation, bond length constraints were applied to all hydrogen-containing bonds using the SHAKE algorithm [24]. The system was equilibrated for 1 ns with a time step of 1 fs. After this period, production dynamics were performed for 4 ns and data were collected every 500 fs. Upon completion of the simulations, analysis of structural parameters was carried out: the C_α root-mean-square deviation (rmsd) for scFv, the OH oxygen rmsd for 1, the dihedral angles associated with the glycosidic linkages in **1** and the intermolecular H-bonds. The geometric criteria used to establish H-bonds are: heavy atom (A) to heavy atom (B) distance (r) $\leq$ 4.0 Å and AHB bond angle $\geq$ 120°. Additionally, the occupancy, i.e., the fraction (f) of the simulation steps for which the H-bond criteria are satisfied, was evaluated.

3.4 Computational Results

To assist in evaluating the structural differences between the (scFv + **1**) complex in solution and in the gas phase, MD simulations were performed on the desolvated (scFv + **1**)$^{n+/-}$ ions at charge states +8 and -8, and several structural parameters were evaluated: C_α rmsd for scFv, the OH oxygen rmsd for **1**, the dihedral angles associated with the glycosidic linkages in **1** and the intermolecular H-bonds.

Uncertainty in the location of the charges is a major challenge to the implementation of MD simulations to large gaseous, multiply charged ions such as proteins and protein complexes. In the present study, ten different charge distributions were considered for the (scFv + **1**)$^{8+}$ and the (scFv + **1**)$^{8-}$ ions. Since the scFv contains eight Arg residues, and because Arg is the most basic amino acid in the gas phase [153], one of the distributions involved protonation of all eight Arg residues. Nine other distributions, in which one or more of the Arg residues were neutralized and the charge was placed instead on Lys or His residues, were also considered. Of the common amino acids, Asp and Glu have the lowest intrinsic gas phase acidities and are, in the absence of other effects, the most likely sites of deprotonation for the negatively charged (scFv + **1**)$^{n-}$ ion [154]. Consequently, the ten charge distributions considered for the (scFv + **1**)$^{8-}$ ion involved deprotonation of Asp and Glu residues. The energies of the (scFv + 1)$^{8+/-}$ ions, at each of the charge distributions considered, were minimized and the charge distribution of the complex was determined by analysing the relative energies of ten different charge distributions. The lowest energy charge distributions were used for the MD simulations.

The C_α rmsd was calculated with respect to the crystal structure of the (scFv + **1**) complex [13]. Values of 1.96 (standard deviation 0.09) and 2.77 (0.22) Å were determined for the (scFv + **1**)$^{8+}$ and (scFv + **1**)$^{8-}$ ions, respectively. Smaller C_α rmsd values were obtained when only the amino acid residues located in vicinity of the ligand binding site were considered, 0.54 (0.07) for +8 and 0.99 (0.12) Åfor -8. Values of 1.77 (0.11) and 1.43 (0.18) Å were calculated for the OH oxygens in **1** at the +8 and -8 charge states, respectively. The small rmsd values, which are comparable in

size to values recently reported by Patriksson and coworkers [134], suggest relatively minor structural changes accompanying the transfer of the complex from solution to the gas phase, at least at the +8 and -8 charge states. The average glycosidic dihedral angles in **1** [ϕ_1($O5^{Gal}$-$C1^{Gal}$-$O2^{Man}$-$C2^{Man}$) and ψ_1($C1^{Gal}$-$O2^{Man}$-$C2^{Man}$-$C3^{Man}$); ϕ_2($O5^{Abe}$-$C1^{Abe}$-$O3^{Man}$-$C3^{Man}$) and ψ_2($C1^{Abe}$-$O3^{Man}$-$C3^{Man}$-$C4^{Man}$)] are for +8: ϕ_1 = 90 (8), ψ_1 = 97 (7), ϕ_2 = 69 (10) and ψ_2 = 92 (10), and for -8: ϕ_1 = 75 (9), ψ_1 = 118 (11), ϕ_2 = 54 (9) and ψ_2 = 100 (9). According to the crystal structure, these angles are: ϕ_1 = 77, ψ_1 =144, ϕ_2 = 72 and ψ_2 = 104 ($\pm$ 10°). This analysis suggests that, at -8, the conformation of **1** is similar to the bioactive conformation in solution. However, at +8, changes in conformation, particularly for the Gal-Man residues, are predicted.

Analysis of the MD trajectories obtained for the (scFv + **1**)$^{8+}$ and the (scFv + **1**)$^{8-}$ ions revealed two general types of intermolecular H-bonds: type 1 interactions, which exhibit a narrow distribution of bond lengths (r) centered at short r ($\sim$ 3 Å), a narrow distribution of bond angles (θ) centered at $\theta > 150°$ and a high occupancy (f > 0.90), and type 2 interactions, which exhibit a broader distribution of r centered at 3.1 to 3.5 Å, a broader distribution of angles, sometimes slightly bimodal in nature and centered at lower values, $\theta < 150°$, and lower occupancy. Although the energies of the identified H-bonds can not be assessed quantitatively, the characteristics of the type 1 interactions are generally associated with strong H-bonds, while the type 2 interactions correspond to weak H-bonds. The distribution of bond lengths and angles determined for the H-bond donor/acceptor pair Man C4 OH/His^{101H}, a type 1 interaction, and the Abe C4 OH/Tyr^{103H} interaction, an example of a type 2 interaction, found for the (scFv + **1**)$^{8+}$ ion are shown in Figure 3.5. The Man C4 OH/His^{101H} interaction persists throughout the simulation (f = 0.97 occupancy) and the distributions of hydrogen bond lengths and angles are narrow, with maxima close to the optimal values (2.86 Å, 162°). In contrast, the Abe C4 OH/Tyr^{103H} interaction has a much lower occupancy (f = 0.51), a markedly larger average r (3.31 Å) and a lower average θ (128°) values. All of the type 1 and type 2 H-bonds, along with the corresponding r, θ and f values are summarized in Table 3.3. The corresponding the

H-bond maps are shown in Figure 3.6.

Eight intermolecular H-bonds were identified for the (scFv + **1**)$^{8+}$ ion from the MD simulations. Of these, five are of the strong, type 1 variety: Trp^{98L}/Abe C4 OH, Gly^{102H}/Abe C2 OH, Trp^{93L}/Gal C4 OH, Gal C6 OH/Asn^{96L} and Man C4 OH/His^{101H}. The weaker, type 2 interactions identified are: His^{34L}/Gal C2 OH, Abe C4 OH/Tyr^{103H}, His^{35H}/Abe C4 OH. Nine H-bonds were identified for the (scFv + **1**)$^{8-}$ ion. Of these, only Gal C4 OH/Asn^{95L} is type 1; the remaining interactions fail to meet one or more of the criteria for a strong H-bond: Trp^{98L}/Abe C4 OH, Gly^{102H}/Abe C2 OH, Gal C6 OH/Asn^{96L}, and Man C4 OH/His^{101H}, His^{34L}/Gal C2 OH, His^{35H}/Abe C4 OH, Abe C4 OH/Gly^{100H} and Trp^{93L}/Gal O (ring). Overall, there is a high degree of structural similarity in the (scFv + **1**)$^{8+/-}$ ions, with six common H-bonds identified. However, important differences are also evident. Notably, the weak Abe C4 OH/Tyr^{103H} interaction found in the (scFv + **1**)$^{8+}$ ion is replaced by a stronger Abe C4 OH/Gly^{100H} interaction in the (scFv + **1**)$^{8-}$. There are also interactions involving Gal C4 OH and Gal O (ring) which are present in the (scFv + **1**)$^{8-}$ ion but absent in the (scFv + **1**)$^{8+}$ ion. Importantly, with the exception of the His^{34L}/Gal C2 OH interaction, all of the conserved H-bonds in the (scFv + **1**)$^{8-}$ ion have larger average r and smaller average θ values than the corresponding interactions in the (scFv + **1**)$^{8+}$ ion. Also, of the three new H-bonds found in the (scFv + **1**)$^{8-}$ ion, only one of these is a type 1 interaction. Therefore, despite the greater number of interactions identified for the (scFv + **1**)$^{8-}$ ion, compared to the (scFv + **1**)$^{8+}$ ion, the individual interactions are likely weaker. Although it is not possible to draw firm conclusions regarding the relative stability of the (scFv + **1**)$^{8+/-}$ ions from the MD data, the present analysis suggests that, despite the greater number of identified interactions, the (scFv + **1**)$^{8-}$ ion is energetically less stable than the (scFv + **1**)$^{8+}$ ion. This prediction is consistent with the lower dissociation E_a determined for the (scFv + **1**)$^{8-}$ ion, compared to the (scFv + **1**)$^{+8}$ ion (See Table 3.1).

Comparison of the H-bond maps predicted by the BIRD/FGR data (see Figure 3.4) and by the MD simulations (see Figure 3.6) affords an opportunity to test of the predictive value of the MD simulations. For the (scFv + **1**)$^{8+}$ ion, the agreement

between experiment and theory is reasonably good. Interactions between Man C4 OH and His^{101H} and between Asn^{96L} and Gal C6 OH were identified with both methods. An interaction between His^{34L} and Gal C3 OH was predicted by BIRD/FGR; according to the simulations, His^{34L} interacts with the neighbouring Gal C2 OH. Also, both methods predict an interaction at His^{35L}, although the binding partner, Gal C4 OH, suggested by simulation could not be confirmed experimentally (due to the unavailability of the corresponding monodeoxy analog of **1**). The BIRD/FGR results suggest energetically important interactions involving, separately, His^{97L} and Man C6 OH. According to the MD simulations, neither the amino acid residue nor the OH group engage in intermolecular H-bonds. The agreement between experiment and theory is less favourable in the case of the $(\text{scFv} + \mathbf{1})^{8-}$ ion. Of the five interactions suggested from the simulations and amenable to experimental testing (Gal C4 OH/Asn^{95L}, Gal C6 OH/Asn^{96L}, Man C4 OH/His^{101H}, His^{34L}/Gal C2 OH, and His^{35H}/Abe C4 OH), only one, the Man C4 OH/His^{101H} interaction, was identified by BIRD/FGR. Based on the results of the above comparison, it is concluded that the MD simulation method, as implemented in the present study, can be used to identify intermolecular interactions within noncovalent biological complexes. However, the method can lead to false positives and false negatives. These shortcomings may reflect limitations in the theoretical model, in particular deficiencies in the choice and use of fixed atomic charges and the choice of charge distributions, as well as the disparity between the computational (ns) and experimental (s) timescales. Additionally, from trajectory analysis it is not possible to quantify the H-bonds and some of the interactions suggested computationally may not be sufficiently strong (> 2 kcal/mol) to be detected using the BIRD/FGR method.

Comparison of intermolecular interactions identified in the gas phase and in the crystal structure Comparison of the intermolecular H-bond maps for the gaseous $(\text{scFv} + \mathbf{1})^{8+/-}$ ions (see Figure 3.4) and the crystal structure of the (scFv + **1**) complex (see Figure 3.2) allows for several conclusions to be drawn. First, there is evidence for the retention of specific H-bonds: His^{101H}/Man C4 OH and His^{34L}/Gal C3 OH in the $(\text{scFv} + \mathbf{1})^{8+}$ ion, and His^{101H}/Man C4 OH in the

(scFv + **1**)$^{8-}$ ion. This is an important finding as it suggests that the structure of the binding site in the (scFv + **1**) complex is at least partially conserved upon transfer of the complex from solution to the gas phase by ES. Secondly, nonspecific intermolecular interactions (i.e., interactions formed in the gas phase, but not present in solution) can play a significant role in stabilizing the protonated (scFv + **1**)$^{8+}$ ion. For example, Gal C6 OH/Asn^{96L} is found in the (scFv + **1**)$^{8+}$ ion from BIRD/FGR and MD simulations (see Figure 3.4 and 3.6). According to the crystal structure of the (scFv + **1**) complex this C6 OH group of Gal is exposed to solvent and, according to microcalorimetry measurements, they contribute little to the binding free energy in solution [155]. Surprisingly, nonspecific intermolecular interactions within the (scFv + **1**)$^{8-}$ ions were not detected. Finally, differences in intermolecular interactions identified experimentally and computationally notwithstanding, the MD data suggest that the loss of the structural water (Wat1) does not result in a dramatic change in the structure of the ligand binding site. Instead, the void created by the loss of Wat1 is filled by new intermolecular interactions between scFv and the Abe residue. Specifically, two of the three residues (His^{35H}, Gly^{100H} and Tyr^{103H}) that are suggested by the crystal structure to interact with Wat1 form new H-bonds with Abe C4 OH: His^{35H}/Abe C4 OH and Abe C4 OH/Tyr^{103H} interactions in the (scFv + **1**)$^{8+}$ ion, and His^{35H}/Abe C4 OH and Abe C4 OH/Gly^{100H} interactions in the (scFv + **1**)$^{8-}$ ion. The behaviour of crystallographic water molecules will be addressed in Chapter 4.

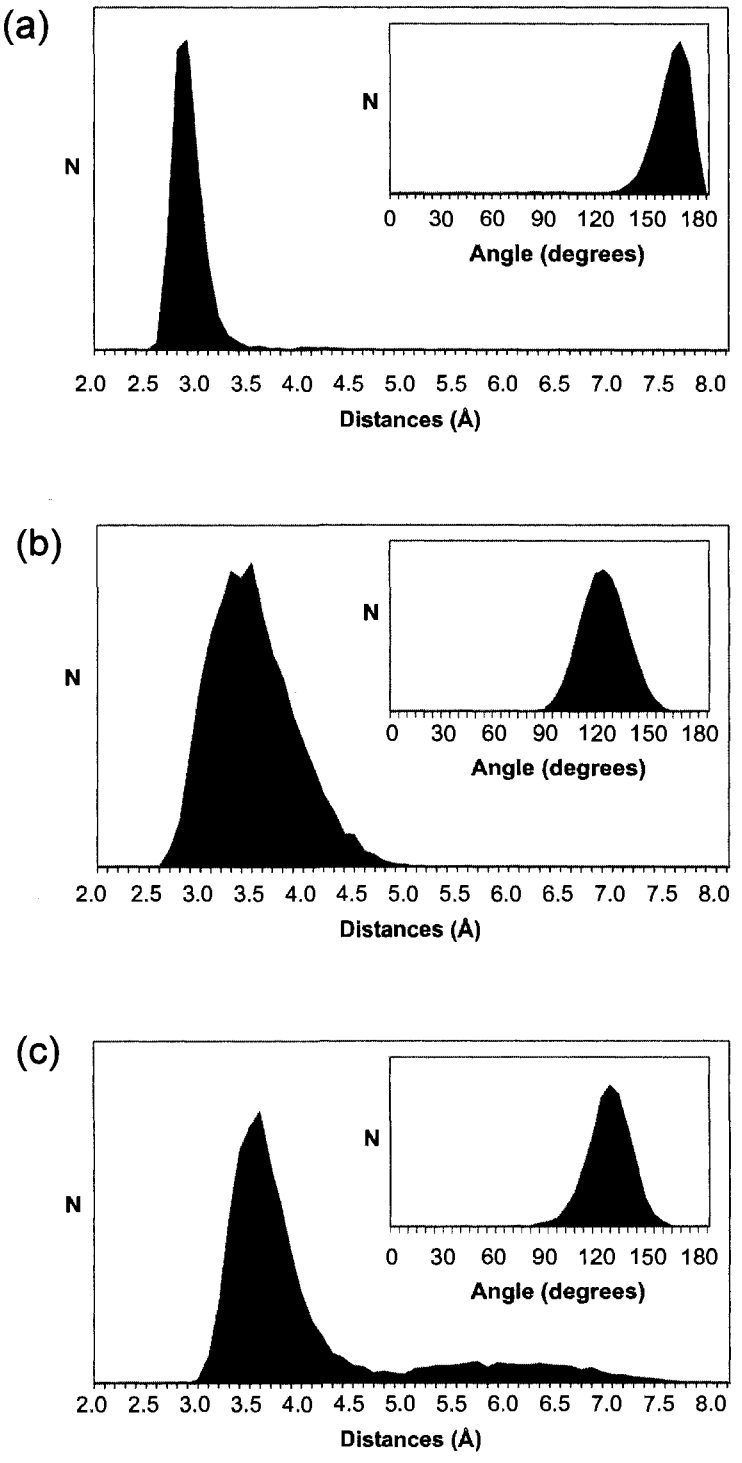


Figure 3.5: Number of occurrences (N) of H-bond distances, and angles (inset), obtained by MD simulations for the (a) Man C4 OH/His^{101H} interaction (type 1) in the (scFV + **1**)$^{8+}$ ion; (b) Abe C4 OH/Tyr^{103H} interaction (type 2) in the (scFV + **1**)$^{8+}$ ion; (c) Man C4 OH/His^{101H} interaction (type 2) in the (scFV + **1**)$^{8-}$ ion.

Table 3.3: Average lengths (r), angles (θ), and occupancy (f) for intermolecular H-bonds within the $(\text{scFv} + \mathbf{1})^{8+}$ and $(\text{scFv} + \mathbf{1})^{8-}$ ions identified from MD simulations.

	$(\text{scFv} + \mathbf{1})^{8+}$			$(\text{scFv} + \mathbf{1})^{8-}$		
H-bond donor/acceptor pair[a] [b]	f	r (Å)	θ (deg)	f	r (Å)	θ (deg)
Trp^{98L}/Abe C4 OH	0.99	2.98 (0.16)	157.69 (9.45)	0.97	3.25 (0.22)	148.30 (10.47)
Gly^{102H}/Abe C2 OH	0.99	2.99 (0.15)	159.21 (10.51)	0.84	3.28 (0.22)	143.14 (12.67)
Trp^{98L}/Gal C4 OH	0.99	2.98 (0.16)	155.97 (10.74)		ND[c]	
Gal C6 OH/Asn^{96L}	0.99	2.77 (0.16)	158.94 (11.02)	0.89	3.06 (0.23)	154.39 (11.93)
Man C4 OH/His^{101H}	0.96	2.86 (0.15)	162.42 (9.71)	0.44	3.54 (0.22)	130.03 (7.25)
His^{34L}/Gal C2 OH	0.72	3.02 (0.19)	146.33 (14.31)	0.94	3.24 (0.23)	158.27 (11.17)
Abe C4 OH/Tyr^{103H}	0.51	3.31 (0.32)	131.74 (8.46)		ND	
His^{35H}/Abe C4 OH	0.18	3.07 (0.18)	127.74 (6.41)	0.24	3.28 (0.24)	132.22 (10.51)
Abe C4 OH/Gly^{100H}		ND		0.92	3.26 (0.25)	153.01 (13.15)
Gal C4 OH/Asn^{95L}		ND		0.90	2.93 (0.16)	156.81 (11.81)
Trp^{93L}/Gal O (ring)		ND		0.86	3.19 (0.22)	134.30 (8.92)

[a]The hydrogen bond distance is given with respect to the heavy atoms.
[b]Values in parentheses correspond to one standard deviation.
[c]ND≡no interaction detected.

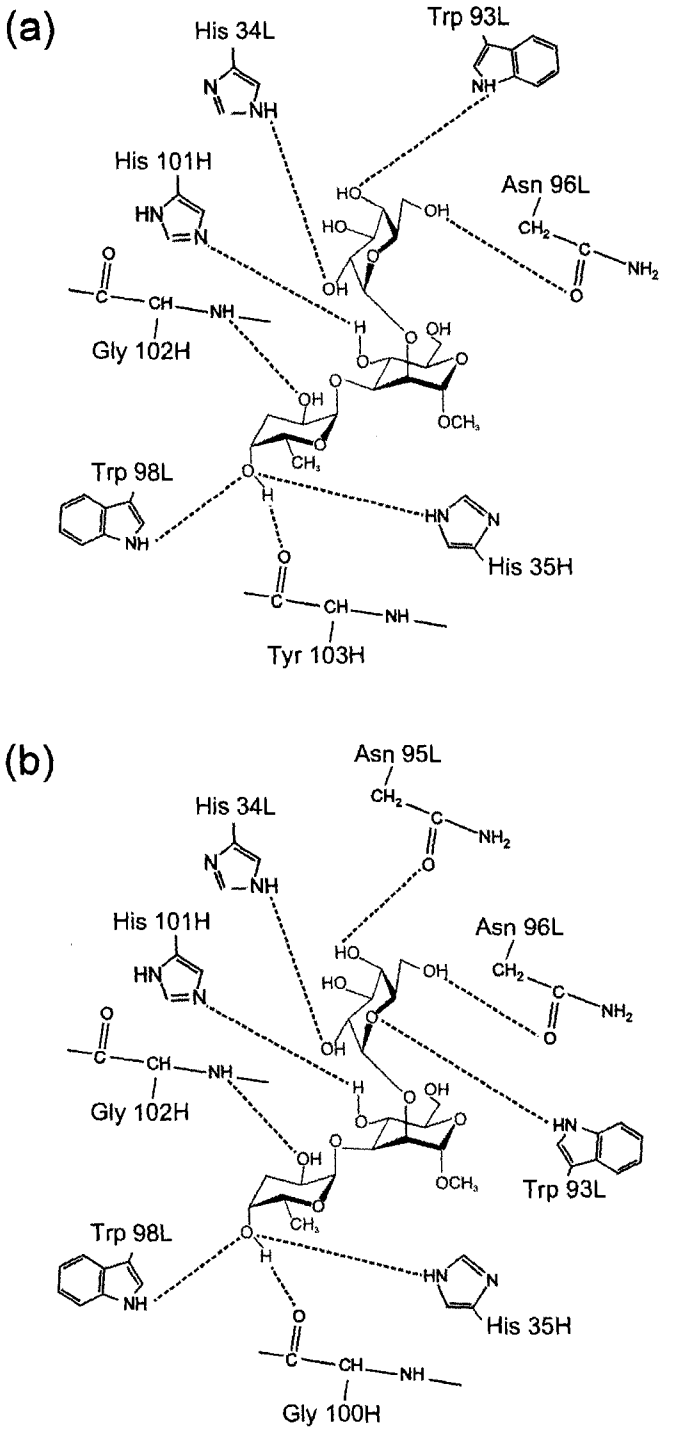


Figure 3.6: Interaction maps determined from MD simulations performed on (scFV + 1)$^{n+/-}$ ions at charge state (a) +8 and (b) -8.

3.5 Conclusions

In conclusion, a detailed study of the intermolecular H-bonds within the protonated and deprotonated ions of a desolvated protein-ligand complex is reported. Using the BIRD/FGR method, we are able to identify intermolecular H-bonds that stabilize the protonated and deprotonated $(\text{scFv} + \mathbf{1})^{n+/-}$ ions. H-bond donor/acceptor pairs are: three pairs (Man C4 OH/$\text{His}^{101\text{H}}$, $\text{His}^{34\text{L}}$/Gal C3 OH and Gal C6 OH/$\text{Asn}^{96\text{L}}$) within the $(\text{scFv} + \mathbf{1})^{8+}$ ions and one (Man C4 OH/$\text{His}^{101\text{H}}$) within the $(\text{scFv} + \mathbf{1})^{8-}$ ions. It is worth noting that two of the above interactions (Man C4 OH/$\text{His}^{101\text{H}}$ and $\text{His}^{34\text{L}}$/Gal C3 OH) correspond to specific intermolecular H-bonds in solution. This strongly suggests that the binding site is, at least partially, conserved upon transfer of the (scFv + **1**) complex from solution to the gas phase by ES. Additionally, other interacting sites on the scFv and on **1**, for which the binding partner could not be elucidated, were identified, as well as nonspecific intermolecular interactions that form upon desolvation. Intermolecular H-bonds were also recognized from MD simulations performed at the +8 and -8 charge states. Our simulations for the $(\text{scFv} + \mathbf{1})^{8+}$ ion showed good agreement with the BIRD/FGR results in predicting a majority of intermolecular interactions; although the agreement was less favourable in the case of the $(\text{scFv} + \mathbf{1})^{8-}$ ion. The structure of the complex at +8 and -8 charge states was found to be different in both the computational and experimental work. In addition to the above results, the computational work also indicated that the nature and strength of the intermolecular interactions can vary with charge state. It was shown that the intermolecular interactions within the $(\text{scFv} + \mathbf{1})^{8-}$ ion are inherently weaker than those within the $(\text{scFv} + \mathbf{1})^{8+}$ ion. Finally, it was suggested from our MD simulations that the water mediated H-bonds between the scFv and **1**, which have been identified in the crystal structure but lost upon transfer of the complex from solution to the gas phase, were replaced with direct H-bonds between **1** and two of the three scFv residues that were originally interacting with the structural water molecule.

Chapter 4

Water Dynamics in Charged and Hydrated Protein-Ligand Complexes

4.1 Introduction

Noncovalent biomolecular complexes between proteins, carbohydrates, small molecules, DNA, and RNA play a key role in many important biological processes [156, 157, 158]. Many biological processes involve the formation and dissociation of specific, noncovalent complexes between proteins and ligands, with solvent molecules playing a significant role in the biophysics of the processes.

However, an understanding of the structure and stability of these complexes and the structural and energetic role played by solvent molecules in the recognition process is incomplete [159]. To achieve a more complete understanding of the molecular recognition process, one must face the challenge of separating solvent effects from intrinsic (solute-solute) interactions. Comparing the structure and stability of biological complexes in solution and in the gas phase is a promising approach to understand the contribution of solvent effects and intrinsic interactions to the binding affinity of protein-ligand complexes.

Specific, noncovalent biological complexes in solution can be converted to the corresponding desolvated gas phase complex through electrospray ionization (ES) [117] and subsequently analysed with a variety of mass spectrometry (MS)-based

techniques. The loss of water molecules which occurs during the ionization process is due to the relatively low strength of the interactions between individual water molecules and the complexes.

Mass spectrometry combined with electrospray is a powerful tool for studying noncovalent biological interactions in the gas phase [149, 117] and ES/MS is also increasingly being used for a variety of applications: measuring the relative binding affinities [142, 143], establishing the composition of biomolecular complexes[160, 161] and investigating the intrinsic interactions in biological complexes [162, 129, 151, 163]. In earlier experimental studies, Klassen and co-workers [129, 151, 163, 128] used the blackbody infrared radiative dissociation (BIRD) technique implemented with a Fourier-transform ion cyclotron resonance mass spectrometer (FT-ICR MS) combined with functional group replacement (FGR) to identify individual interactions in gaseous biological complexes and to quantify the strength of the interactions.

The biological complex of interest in the current study is composed of a single chain variable fragment (scFv) of the monoclonal antibody Se155-4 and its natural trisaccharide ligand, α-D-Gal$p$(1$\rightarrow$2)[α-D-Abe$p$(1$\rightarrow$3)]-α-D-Man$p$(1$\rightarrow$OMe) (**1**). Its crystal structure and that of the corresponding antigen binding fragment (Fab) complex have been solved [13]. Analysis of the crystal structures suggests that a trisaccharide ligand is bound to the scFv through intermolecular hydrogen bonds in solution. Additionally, water molecules belonging to a network of well-ordered solvent molecules surrounding the trisaccharide in the scFv structure were observed. These "crystallographic" water molecules participate in H-bonds with both the ligand and the protein. In the crystal structure, the first water (Wat 1) is located in the base of the binding pocket, and mediates H-bonds between the scFv protein and the ligand **1**. Two additional waters (Wat 2, Wat 3) are also observed in the crystal structure (see Figure 4.1).

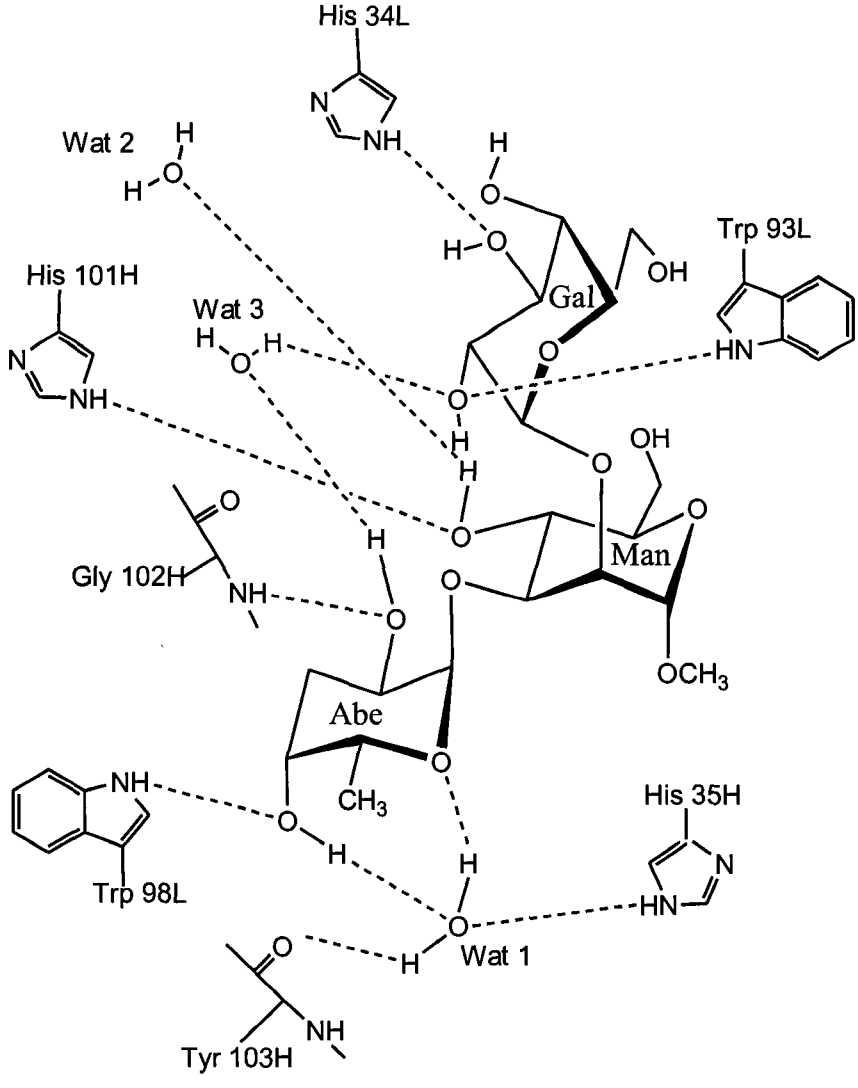


Figure 4.1: Intermolecular hydrogen bond scheme for the (scFv + **1**) complex obtained from X-ray analysis of the crystal structure.

Recently,[1] we have presented the detailed and quantitative investigations of the intermolecular interactions within gaseous ions of the desolvated complex described above over a range of charge states using both experimental and computational methods [162]. Our analysis identified H-bonds between the protein and the ligand in different charge states and suggested that there was a conservation of the binding site, to some extent at least, upon transfer of the complex from solution to the gas phase. In the same study, we compared the noncovalent interactions identified in the gas-phase with the aid of computer simulations with those present in the crystal structure. It was shown that water-mediated H-bonds between the protein and ligand, which are originally present in the crystal structure and lost upon transfer of the complex from solution to the gas phase, are replaced with direct H-bonds between the protein and the ligand.

It is of great interest to extend our study to investigate the dynamics of "crystallographic" water molecules in the charged and hydrated protein-ligand complex in the gas phase. As mentioned above, the hydration waters are rapidly lost in the gas phase and the "crystallographic" water molecules are the last ones to evaporate during the desolvation process because they participate in H-bonds with the complex (see Figure 4.1). Thus, MD simulations are performed in the gas phase including only those "crystallographic" water molecules. Here, we performed a series of MD simulations from low to high temperatures in order to study the temperature dependence of the individual water dynamics around the binding site of the charged protein-ligand complex in the gas phase. As opposed to our previous efforts [162], where equilibrium properties were investigated in the form of the identification of H-bonded interactions, the current study probes the dynamics of water molecules and utilizes the concept of hydrogen bond lifetime to determine the structural relaxation of H-bonds between water molecules and the complex.

The dynamics of water molecules can be experimentally measured by dielectric relaxation [164], NMR spectroscopy [165], solution X-ray, and neutron scattering [166]. A number of experimental techniques have probed properties of hydration wa-

[1]See Chapter 3

ter molecules on time scales between microseconds and nanoseconds [167]. Recently, terahertz absorption spectroscopy was applied as a probe for the fast solvation dynamics around a chosen solute, lactose [167]. This approach was used to detect the solute-induced changes in the water network near lactose, where the H-bond rearrangement dynamics of water molecules occurs on the picosecond time scale. The dynamics of water molecules can be observed over a broad range of time scales in different physical phenomena, but a complete picture of the nontrivial interactions of hydration water molecules is still lacking [168].

In recent years, simulation techniques have proven to be an increasingly powerful tool in modelling the behaviour of water molecules [169]. With such simulations, we can track individual water molecules at the molecular level, something currently beyond the capability of experiments. There has been a number of simulation studies, primarily focused on the relaxation behaviour of H-bonds in pure water [170, 171, 172] as well as in aqueous solutions of electrolytes and micelles [173]. The concept of hydrogen bond lifetime has been used to study H-bond dynamics for carbohydrates in solution [174, 175], in an aqueous micelle [176, 177], and recently in DNA groove water dynamics [178, 179, 180].

In the present study, we aim at providing a direct microscopic picture in terms of H-bond dynamics of the individual water molecules at different temperatures in the charged and hydrated protein-ligand complex in the gas phase. The hydrogen bond lifetimes were obtained for H-bonds between water molecules and the complex. Note that this lifetime is on the picosecond time scale, and corresponds to the forming and breaking of protein-water or ligand-water H-bonds.

4.2 Computational Methods

The AMBER 9.0 program suite [109] was used for the molecular dynamics simulations. The initial geometries for a single chain variable fragment (scFv) of the monoclonal antibody Se155-4 and a trisaccharide ligand, αGal[αAbe]αMan(**1**) were taken from the Brookhaven protein database (pdbid=1MFA) [13]. The simulations were

performed using the AMBER 94 forcefield [47] with the GLYCAM [12] parameter set for the ligand. We conducted simulations for the +8 charge state of the protein. The details of the chosen charge distribution of the protein in this study are reported elsewhere [162]. Electrostatic potential (ESP) atomic partial charges, determined by Woods *et* al., [60] were used for the ligand. Three "crystallographic" water molecules, described by the TIP3P model [110], were included in the system. A series of MD simulations were performed over a wide temperature range (25, 50, 75, 100, 150, 200, 250, 300, 350, 400, 450 and 500 K). The energy of the system was first minimized with the conjugate gradient method using a 0.0001 kcal/mol Å convergence criterion. The entire system was then gradually heated from 10 K to the desired temperature over a period of 20 ps. Simulations were first performed in vacuum and in the canonical ensemble (NVT). Constant temperature was maintained using the weak-coupling algorithm with time constant of 1.0 ps [152]. The system was equilibrated for 1 ns with a time step of 1 fs. After this period, production dynamics were performed for 4 ns and data was collected every 500 fs. In order to study the dynamics of water molecules, correlation functions were calculated at each temperature from a total of 493 microcanonical (NVE) production runs (150 ps for each) with initial conditions sampled from the initial canonical simulation. During the simulations, bond length constraints were applied to all hydrogen-containing bonds using the SHAKE algorithm [24]. Upon completion of the simulations, we carried out a structural hydrogen bond analysis. The geometric criteria used to establish H-bonds are: heavy atom (A) to heavy atom (B) distance (r) $\leq$ 4.0 Å and AHB bond angle (θ) $\geq$ 120°. Additionally, the occupancy, *i.e.*, the fraction (f) of times that the H-bond criteria are satisfied, was evaluated. For the H-bonds identified using a structural hydrogen bond analysis, the hydrogen bond lifetime correlation functions were calculated.

4.3 Results and Discussion

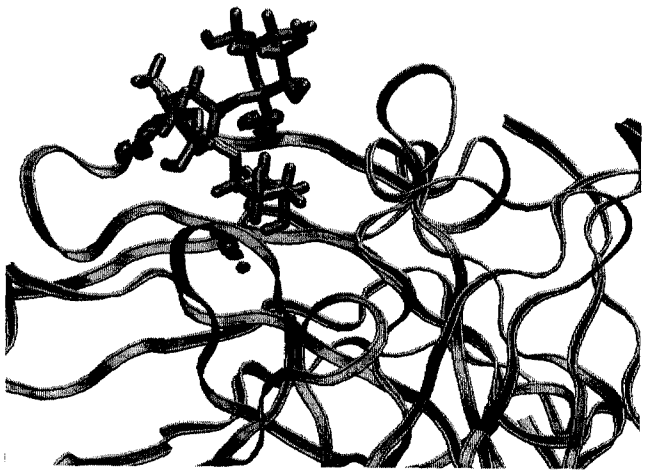
In general, the crystal structure is only representative of a single configuration. At very low temperatures, the configurations explored during the simulation should be

similar to the crystal structure. To illustrate this point, we show the water densities at 25 and 50 K in Figure 4.2. Those were obtained by calculating a three dimensional histogram of the oxygen and hydrogen atoms of all the water molecules over a canonical simulation. In order to obtain a consistent frame of reference, the configuration of the system was adjusted via an RMS fit to the initial (time zero) configuration of the trajectory. The frame of reference of this fit is based on the ligand and residues of the active site. For the purpose of comparison, we used the same density isovalue in all the figures. Oxygen density is represented in red and hydrogen density is in gray. The ligand and the protein are in orange and green, respectively.

As temperature is increased, we observed that the water mobility also increases. This prompts us to investigate the temperature dependence of the dynamical behaviour for individual water molecules. The water densities at different higher temperatures are shown in Figure 4.3. We observed that the most buried water molecule (Wat 1), which interacts with Abe and key amino acid residues, remains localized while the other two become more and more delocalized as the temperature is increased up to 300 K. At this temperature, these latter two water molecules have left the binding site and interact with the surface of the protein.

We have made an initial estimate that Wat 1 can *survive* to temperatures of up to 300 K. Thus, all crystallographic waters leave the complex above 300 K. The analysis of the trajectories was therefore carried out up to 300 K. We chose to report here the simulations results obtained at three temperatures (100, 200 and 300 K) and we now focus on the dynamical behaviour of these crystallographic water molecules.

(a)



(b)

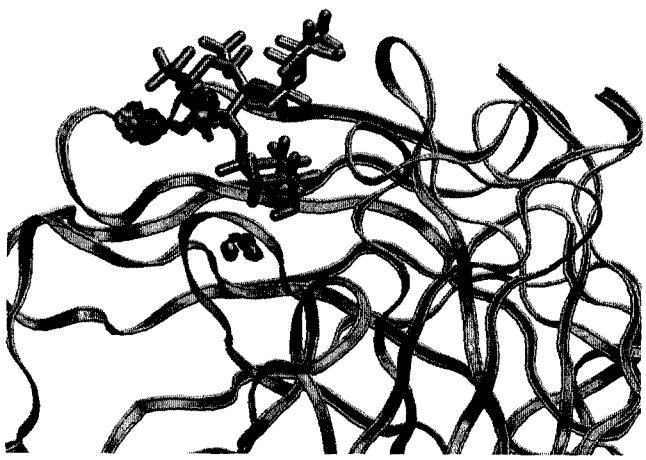


Figure 4.2: Water density at (a) 25 K and (b) 50 K.

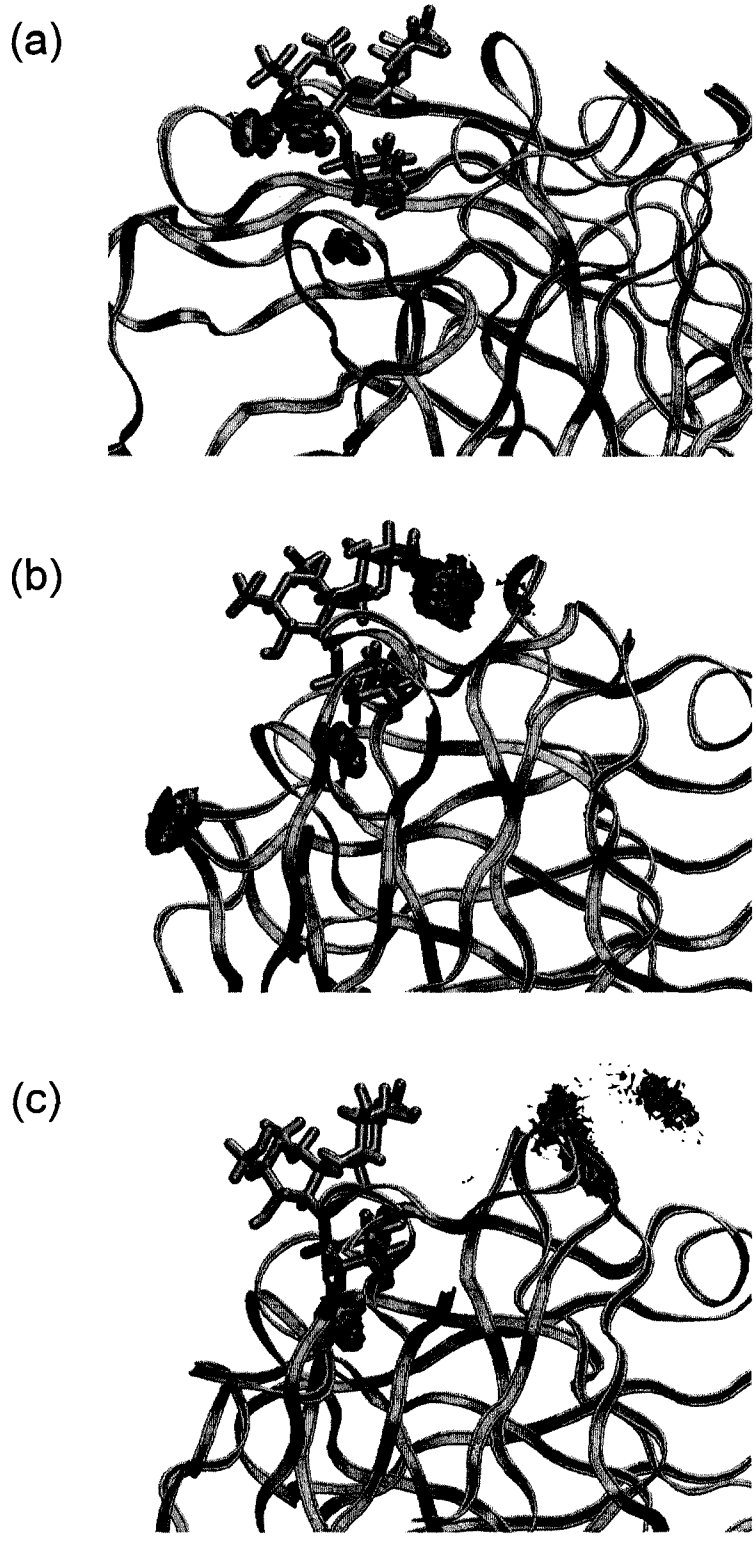


Figure 4.3: Water density at (a) 100 K, (b) 200 K and (c) 300 K.

4.3.1 Hydrogen bond lifetime dynamics

Water molecules can form H-bonds both with the amino acids of the protein and with hydroxyl groups of the ligand. We have studied the protein-water and ligand-water H-bond lifetime dynamics. Table 4.1 shows the occupancy (f) of the identified H-bonds between water molecules and the complex. We observed a competition between the two hydrogens of Wat 1 to form a H-bond with Tyr^{103H} at 200 and 300 K. In our previous study on protein-ligand interactions [162], we showed that the Tyr^{103H} also makes a H-bond as an acceptor with the Abe C4 OH group of the ligand. This Abe C4 OH/Tyr^{103H} interaction corresponds to one of the weak *type 2* (see Section 2.4) identified H-bonds within (scFv + **1**)$^{8+}$ ion in the gas phase. Using the information in Table 4.1, we note that when Wat 1 is acting as a H-bond acceptor, the H-bonds do not show a dynamical behaviour. This is because Wat 1 makes a strong interaction with His^{35H} with a high occupancy (f = 1.00 at 100 and 200 K, and f = 0.96 at 300 K). The high occupancy indicates that the His^{35H}/Wat 1 interaction persists throughout the simulation. A similar behaviour is observed for the Abe C4 OH/Wat 1 interaction. Therefore, these two interactions (*i.e.,* both strong H-bonds of Wat 1 acting as an acceptor with ligand and protein) serves to explain why Wat 1 is locked in the binding site over the range of temperatures studied. Those strong interactions of Wat 1 are in agreement with the crystal structure (see Figure 4.1).

We found that Wat 2 makes two strong H-bonds (Man C4 OH/Wat 2 and Wat 2/His^{101H}) at 100 K. These H-bonds display narrow distributions of r, θ and a high occupancy. However, no interactions were detected for these pairs at higher temperatures. Similarly, strong H-bonds were also identified for Wat 3: Trp^{33H}/Wat 3 and Wat 3(H1)/Man C4 OH at 100 K and Asn^{55L}/Wat 3 at 200 K. As in the case of Wat 2, these interactions do not exist at 300 K.

Table 4.1: Occupancy (f) for water hydrogen bonds within the (scFv + **1**)$^{8+}$ ion at T = 100, 200 and 300 K.

H-bond donor/acceptor pair	100 K	200 K	300 K
His^{35H}/Wat 1	1.00	1.00	0.96
Abe C4 OH/Wat 1	1.00	0.99	0.83
Wat 1(H1)/Tyr^{103H}		0.18 (biθ)	0.42 (biθ)
Wat 1(H2)/Tyr^{103H}	0.89 (biθ)[a]	0.80 (biθ)	0.50 (biθ)
Wat 1 (H1)/Gly^{99H}	0.81 (biθ)		
Wat 1(H1)/Gly^{100H}	0.10 (biθ)		
Wat 1(H2)/Gly^{100H}	0.25 (biθ)		
Man C4 OH/Wat 2	0.97		
Wat 2/His^{101H}	0.94		
Ser^{94L}/Wat 2		0.72 (bir, biθ)	
Gal C2 OH/Wat 2	0.32 (bir)		
Gln^{1L}(H1) /Wat 2			0.18 (bir, biθ)
Gln^{1L}(H2) /Wat 2			0.24 (bir, biθ)
Gln^{1L}(H3) /Wat 2			0.23 (bir, biθ)
Wat 2(H1)/Gly^{32L}		0.29 (bir)	
Wat 2(H2)/Gly^{32L}		0.29 (bir, biθ)	
Wat 2(H1)/Asn^{95L}			0.25 (bir, biθ)
Wat 2(H2)/Asn^{95L}			0.20 (bir, biθ)
Wat 2(H1)/Gal C4 OH		0.16 (bir, biθ)	
Trp^{33H}/Wat 3	1.00		
Asn^{55L}/Wat 3		0.95	
Wat 3(H1)/Man C4 OH	0.97		
Wat 3(H1)/Asn^{54L}		0.54 (biθ)	
Wat 3(H2)Asn^{54L}		0.43 (biθ)	
Wat 3(H1)/Asn^{95L}			0.18 (bir, biθ)
Wat 3(H2)/Asn^{95L}			0.15 (bir, biθ)
Gln^{1L}(H1) /Wat 3			0.27 (biθ)
Gln^{1L}(H2)/Wat 3			0.23 (biθ)
Gln^{1L}(H3)/Wat 3			0.22 (biθ)

[a]bir $\equiv$ bimodal distribution of r, biθ $\equiv$ bimodal distribution of θ, otherwise narrow distributions of r or θ.

The structural relaxation of H-bonds can be characterized by the hydrogen bond lifetime correlation function (HBLTCF) given by [171, 172, 178, 179, 180]

$$S(t) = \frac{\langle h(0)H(t)\rangle}{\langle h\rangle} , \tag{4.1}$$

$$C(t) = \frac{\langle h(0)h(t)\rangle}{\langle h\rangle} , \tag{4.2}$$

where $h(t)$ and $H(t)$ are the hydrogen bond probability variables. The quantity $h(t)$ is 1 if a particular pair of protein-water or ligand-water, is hydrogen bonded at given time t and 0 otherwise. On the other hand, $H(t)$ is equal to 1 if a particular pair is continuously hydrogen bonded up to time t and 0 otherwise. Therefore, $C(t)$ allows reformation of a bond that is broken at some intermediate time while $S(t)$ decays as soon as the bond breaks for the first time.

The HBLTCF were calculated for H-bonds that show bimodal distributions of distances (r) or angles (θ) with an occupancy $f \geq 0.1$.[2] These particular H-bonds were chosen because of their dynamical behaviour as opposed to stronger H-bonds that have extremely large lifetimes on the time scale of the simulation. Our previous example of a strong and long lived interaction (His^{35H}/Wat 1) was given in Figure 4.4(a). This interaction exhibits a narrow distribution of r centered at a short value ($\sim$3 Å), a narrow distribution of θ centered at $\theta \geq 150°$ and a high occupancy (see Table 4.1).

In earlier studies on various aspects of H-bond lifetime dynamics in bulk water [170, 171, 172] or DNA grooves [180], the average time correlation functions over all the H-bonds were computed to obtain a measure of dynamics. However, each water molecule around the protein-ligand complex in our study behaves differently, so we need to study each H-bond separately. Thus, to find the H-bond geometric criteria used in calculating the HBLTCF, we looked at the distributions of r and θ determined for each selected H-bond pair at each temperature. An example is shown in Figure 4.4(b) where one can identify the minimum between the two peaks of the distributions

[2]Note that the occupancy was obtained using a criterion of $r \leq 4.0$ Å and $\theta \geq 120°$ as stated earlier.

as the cutoff point (criterion for H-bond). For example, the Ser^{94L}/Wat 2 bond is said to exist if r is less than 4.0 Å and θ is greater than 109.4°.

Using the above geometric criteria, we calculated the HBLTCFs for the H-bonds that show their dynamical behaviour. Figure 4.5 contains the HBLTCFs for the Wat 1(H2)/Tyr^{103H} interaction at different temperatures, calculated over a time of 150 ps. Note that to obtain the H-bond criteria, we averaged over the different temperatures in order to have the same criteria at all temperatures. The Wat 1(H2)/Tyr^{103H} interaction shows a strong temperature dependence. The H-bond lifetime dynamics for the Wat 1(H2)/Tyr^{103H} interaction at 300 K is faster than ones at 100 or 200 K. A slow tail in the decay of $S(t)$ is observed at 200 K and it becomes more important at 100 K. It can be seen from Figure 4.5 that the decay behaviour of the $C(t)$ correlation function is much slower. This makes sense since that quantity is associated with intermittent H-bonds.

The characteristic decay time of the continuous hydrogen lifetime correlation function $S(t)$ yields an estimate for the H-bond lifetime (τ_{HB}) of each the protein-water and ligand-water H-bond. We have fitted these $S(t)$s to a single exponential form to get τ_{HB}, and H-bond lifetimes are provided in Table 4.2. Generally, the H-bond lifetime (τ_{HB}) decreases as temperature increases for the same interaction pair. In the case of the Wat 1(H2)/Tyr^{103H} interaction, the H-bond lifetime is much longer at 100 K than at higher temperatures. It was observed that contrary to Wat 1, Wat 2 and Wat 3 show much faster decay of their H-bonds. Also, the temperature dependence of the H-bond lifetime of water H-bonds is large. We see that Wat 2 and Wat 3 diffuse away from the binding site at high temperatures with shorter τ_{HB}. These observations are consistent with the water densities shown in Figure 4.3.

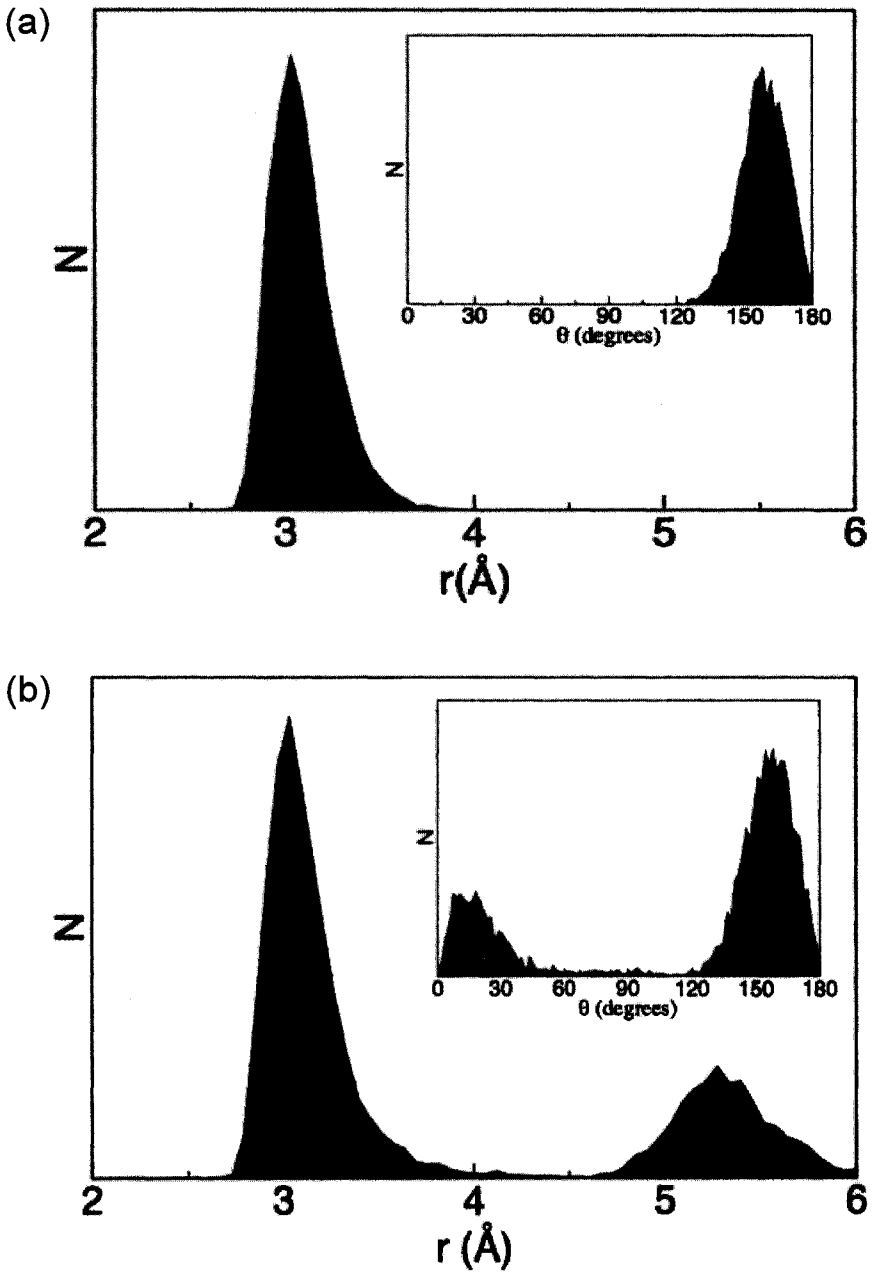


Figure 4.4: Number of occurrences (N) of H-bond distances (r), and angles (θ) (inset) for (a) the His^{35H}/Wat 1 and (b) the Ser^{94L}Wat 2 interactions at 200 K.

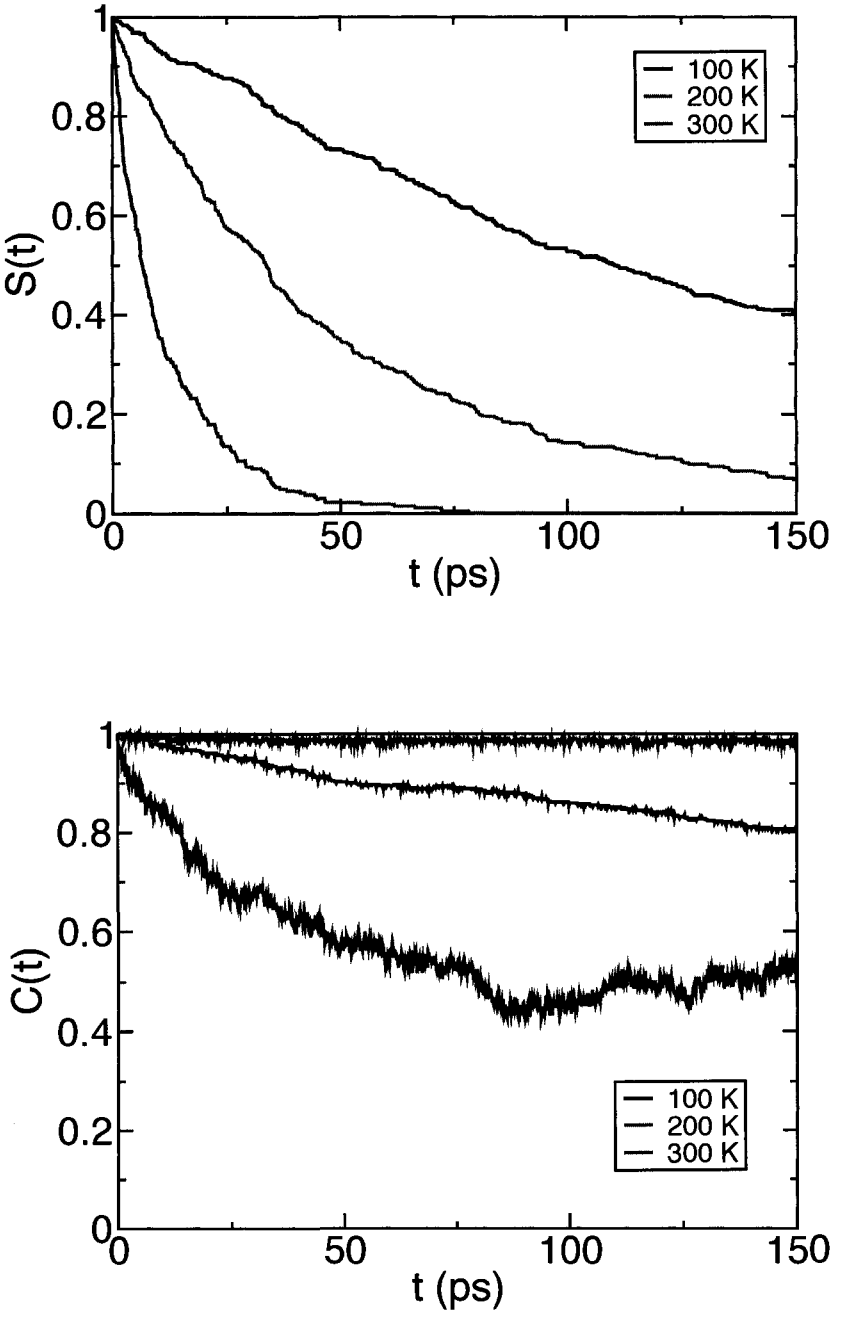


Figure 4.5: Hydrogen bond lifetime correlation functions, (a) $S(t)$ and (b) $C(t)$ for the Wat 1(H2)/Tyr^{103H} interaction at T = 100, 200 and 300 K.

Table 4.2: Hydrogen bond lifetime τ_{HB} (in ps) for water hydrogen bonds at $T = 100$, 200, and 300 K.

H-bond donor/acceptor pair	100 K	200 K	300 K
Wat 1(H1)/Tyr^{103H}		48.0[a]	11.0
Wat 1(H2)/Tyr^{103H}	157.9	52.2	12.6
Wat 1(H1)/Gly^{99H}	10.7		
Wat 1(H1)/Gly^{100H}	1.1		
Wat 1(H2)/Gly^{100H}	11.2		
Ser^{94L}/Wat 2		47.8	
Gln^{1L}(H1)/Wat 2			2.0
Gln^{1L}(H2) /Wat 2			2.2
Gln^{1L}(H3)/Wat 2			2.1
Wat 2(H1)/Gly^{32L}		1.4	
Wat 2(H2)/Gly^{32L}		1.0	
Wat 2(H1)/Asn^{95L}			2.6
Wat 2(H2)/Asn^{95L}			2.1
Wat 2(H1)/Gal C4 OH		2.3	
Wat 3(H1)/Asn^{54L}		35.6	
Wat 3(H2)/Asn^{54L}		27.7	
Wat 3(H1)/Asn^{95L}			4.6
Wat 3(H2)/Asn^{95L}			0.7
Gln^{1L}(H1)/Wat 3			2.4
Gln^{1L}(H2)/Wat 3			3.7
Gln^{1L}(H3)/Wat 3			2.8

[a] τ_{HB} was calculated for H-bonds that show bimodal distributions of r or θ with the occupancy $f \geq 0.1$ in Table 4.1.

4.4 Conclusions

In this chapter, we have explored in detail the H-bond dynamics of three "crystallographic" water molecules in a charged protein-ligand complex by means of MD simulations. These water molecules participate in H-bonds with both the scFv protein and the ligand in the crystal structure. A series of MD simulations were performed from low to high temperatures in order to study the temperature dependence of the individual water dynamics around the binding site of the complex in the gas phase. The calculated water densities suggested that the water mobility also increases as temperature is increased. It was shown that the first water molecule (Wat 1) located in the deep binding pocket remains localized while the other two (Wat 2, Wat 3) become more and more delocalized as the temperature is increased up to 300 K. Using a structural H-bond analysis, intermolecular H-bonds between water molecules and the complex were identified. In order to understand the diversity in the nature of water-protein and water-ligand H-bonds around the protein-ligand complex, we calculated H-bond lifetime correlation functions for the H-bonds that show a dynamical behaviour. Lifetimes were obtained for these H-bonds by exponential fits. The calculations revealed that the H-bond lifetime decreases as temperature increases, and the structural relaxation of the H-bonds formed at high temperatures has been found to be faster than those at low temperatures. It was also observed that the relaxation of H-bonds formed with Wat 1 is much slower than that of those formed with Wat 2 or Wat 3.

Chapter 5

Dissociation Kinetics of Protein-Ligand Complexes

5.1 Introduction

Essential biological processes are dependent upon specific interactions between biological molecules. Thus, understanding how two molecules recognize each other is one of the fundamental and central issues in many biological processes. A number of experimental and theoretical efforts have gone into the study of elucidating the mechanisms involved in recognition processes for these reasons [181, 182, 183, 184, 185, 186]. To understand fully the molecular recognition phenomena in the vast majority of chemical and biological processes, a close examination of the underlying binding free energy behaviour is often necessary.

Free energy is an important thermodynamic property. It is expressed in two forms: the Helmholtz free energy A and the Gibbs free energy G which can be defined as $A = U - TS$ and $G = H - TS$ where U is the internal energy, T is the temperature, S is the entropy and H is the enthalpy. The accurate prediction of binding free energy is one of the challenges in molecular modelling because free energy differences between different molecular states are directly related to experimental observables, from binding affinity or equilibrium constants in the context of equilibrium thermodynamics, to kinetic rate constants via transition state theory. We can obtain important insights and improved understanding of a wide variety of chemical phenomena, such as ligand binding and

mutations, from a accurate knowledge of free energy, especially by breaking down the total free energy change into contributions from solvent, protein and even individual residues or chemical groups. For instance, protein-ligand binding constants are of importance in the emerging field of *de novo*, rational drug design, and cannot be predicted reliably and accurately without the knowledge of the associated free energy changes [37].

The discovery of a ligand that binds a targeted protein with high affinity in the field of rational drug design, while keeping favourable pharmacological properties, is a major and costly challenge [187]. Computer simulations can help circumvent some of the difficulties. Indeed, computer simulation plays a significant role in guiding molecular design due to its potential for predicting accurate protein-ligand binding free energies.

Over the years, a variety of computational methods have been used to determine binding free energies in complex biomolecular systems. One particular class of such methods involves the calculation of the potential of mean force (PMF) along a reaction pathway. This pathway is usually referred to as the reaction coordinate and the PMF is a free energy profile determined along the chosen reaction coordinate [39, 188]. Different techniques have been successfully applied for the calculation of the PMF profile. Most commonly used techniques include free energy perturbation (FEP) [189, 190], thermodynamics integration (TI) [191], and umbrella sampling [192] with the weighted histogram analysis method (WHAM) [42, 193].

In applications to protein-ligand binding, the free energy perturbation (FEP) methodology was first applied to compute relative free energies [189]. Since the first applications of FEP to the calculation of relative free energies were reported, theoretical and computational tools to predict binding free energies with quantitative accuracy were developed and improved, making the free energy simulation techniques a well-characterized modelling tool for drug design [194, 195, 196]. Also, developments on both formal and technical aspects have contributed to decrease the computational cost of free energy calculations.

Recently, the usage of the conformational and orientational restraint potentials

was introduced into the calculation of protein-ligand binding free energy to improve convergence by modelling a flexible ligand as a relatively rigid one [185, 183, 197]. With this approach, the difficulties associated with exploring a multitude of conformations are significantly reduced because the ligand does not have to sample the entire simulation volume [185, 183, 197]. Those restraint potentials also helped to improve the computational efficiency. The choice of restraints for our case of interest will be discussed further in Section 5.2.4.

In the current chapter, we present our work on the free energy simulation study of the dissociation kinetics of a series of structurally related noncovalent protein-ligand complexes in the gas phase. The protonated ion of a complex composed of a single chain fragment (scFv) of a monoclonal antibody and its native trisaccharide ligand, αGal[αAbe]αMan (**1**) was chosen for this study. A series of single amino acid mutants, single ligand mutants and double amino acid-ligand mutants were also chosen. Details of the studied systems will be provided in Section 5.3. Through simulations of various mutants, the free energy change associated with substituting a chemical group with another is evaluated. This allows us to predict such effects as the influence of point mutations on thermal stability and ligand binding of proteins, or the role of different substituents in determining the affinity of the ligand for a protein.

A number of experimental studies on scFv-oligosaccharide have been reported. The association thermodynamics for Se155-4 with a variety of oligosacchrides has been extensively investigated in solution [198]. Recently, Klassen and co-workers developed a reactivity-based approach, employing blackbody infrared radiative dissociation (BIRD) [126, 127, 199, 200], a thermal dissociation technique implemented with a Fourier-transform ion cyclotron resonance mass spectrometer (FT-ICR MS). They reported the first time-resolved thermal dissociation kinetic and energetic measurements, carried out using the BIRD technique, for a series of scFv-trisaccharide complexes [128]. In this study, BIRD was used to measure the thermal dissociation rate constants (k) and to determine Arrhenius activation parameters for the dissociation of noncovalent protein-ligand complexes.

The PMFs for the dissociation of protein-ligand complexes were obtained using

MD simulations. Restraint potentials were applied into the calculation of the protein-ligand binding free energy to solve the convergence problem which arises from the flexibility of the unbound ligand [183, 185]. Technical details regarding the systems and the computational procedure are given in Section 5.3. The dissociation rate constants for the desolvated protein-ligand complexes can be computed from the PMFs by employing the variational transition state theory (VTST) [201, 202, 201, 203]. Further discussion of the application of PMF to kinetics as well as the definition of PMF will be described in Section 5.2. Here, our goal is to present detailed computational studies aimed at calculating PMFs for the dissociation of protein-ligand complexes in the gas phase, and computing dissociation rate constants from the free energy of the complexes.

5.2 Theory

5.2.1 Potential of mean force and umbrella sampling

The protein-trisaccharide complex is a reactant in the process leading to its dissociation into two components, the protein and trisaccharide. The resulting PMF $\omega(r)$ along some coordinate r is defined from the average distribution function $\langle\rho(r)\rangle$ [41]:

$$w(r) = w(r^*) - k_B T \ln\Big[\frac{\langle\rho(r)\rangle}{\langle\rho(r^*)\rangle}\Big] \,, \tag{5.1}$$

where k_B and T are the Boltzmann constant and temperature, respectively. $w(r^*)$ and $\langle\rho(r^*)\rangle$ are arbitrary reference values. Since only differences in PMF are used in calculating rate constants, the choice of $w(r^*)$ does not affect the results. The details of the PMF formulation and the umbrella sampling method were given in Section 1.4.5.

5.2.2 Variational transition state theory: application of PMF to kinetics

Transition state theory (TST) [204, 205, 206, 201] plays a central role in calculating rates of chemical reactions occurring in the gas phase, in condensed phase or in

enzymes. Classically, the fundamental assumption of TST is that there exists a hypersurface (or surface) in phase space which divides space into a reactant region and a product region. TST provides the equilibrium rate constant from the one-way flux through this dividing surface. The classical reactive flux across a dividing surface in a given direction is greater than or equal to the exact classical reactive flux. This upper bound to the exact classical, equilibrium rate constant is the basis for classical variational transition state theory (VTST) in which the best estimate of the rate constant can be obtained by variationally optimizing the dividing surface. Therefore, the transition state is represented by a dividing surface, corresponding to the lowest upper bound, *i.e.* a minimum in the reactive flux [207, 208].

By employing the TST, an upper bound to the dissociation rate constants for the desolvated protein-ligand complexes is given by [201, 202]

$$k_{r^\dagger}^{TST} = -\sqrt{\frac{1}{2\pi m k_B T}} \left(\frac{dA(r)}{dr}\right)_{r^\dagger} , \tag{5.2}$$

where m is the reduced mass of the complex and $A(r)$ is Helmholtz free energy. The quantity r is the reaction coordinate and the dividing surface $r^\dagger$ separates the associated complex (reactant) and dissociated complex (product) phases. The reactant and product regions are defined as $r < r^\dagger$ and $r > r^\dagger$, respectively. The PMF is calculated directly from computer simulations and $A(r)$ can be expressed in terms of the PMF [202]:

$$A(r) - A(r_0) = -k_B T \ln \left[\frac{\gamma}{2!} \int_{r_0}^{r} \exp\left(-\frac{w(r)}{k_B T}\right) r^2 dr\right] , \tag{5.3}$$

where $\gamma = (2\pi m k_B T/h^2)^{3/2}$ and h is the Planck constant. $A(r_0)$ is the reference Helmholtz free energy of a complex at a specific r_0. Eq. 5.4 is the differentiated form of $A(r)$ and it is proportional to the TST dissociation rate constant in Eq. 5.2,

$$-\left(\frac{dA(r)}{dr}\right)_{r^\dagger} = k_B T \frac{\exp\left(-\frac{w(r^\dagger)}{k_B T}\right) r^{\dagger 2}}{\int_{r_0}^{r^\dagger} \exp\left(-\frac{w(r)}{k_B T}\right) r^2 dr} . \tag{5.4}$$

To obtain a minimum in reactive flux according to VTST, the value of dividing surface $r^\dagger$ should be determined to minimize $-\left(\frac{dA(r)}{dr}\right)_{r^\dagger}$ because the reactive flux based on TST is always greater than the real flux.

5.2.3 Reaction coordinates

The PMF in Eq. 5.1 depends on the choice of the reaction coordinates, r. The improper choice of a reaction coordinate can cause the simulation bias and yield slower convergence [209]. Especially for complex systems, it is not always clear how to best determine the reaction coordinate. Geometrical parameters such as a distance, a dihedral angle or a torsion are widely used as a simple function in most work [210, 211]. For example, the distance $r_{AB} = |\mathbf{r}_A - \mathbf{r}_B|$ between two atoms or monatomic ions A and B has been used to study the dissociation of ion pairs [212] and enzyme-catalysed reactions [213]. A similar results is obtained when the reaction coordinate is the distance between the center-of-mass (COM) of two molecules or molecular fragments or between an atom and the COM of a collection of atoms. These types of reaction coordinates are called Jacobi distances [214].

The use of a geometrical variable as the reaction coordinate is particularly instructive and intuitive for chemists and biochemists to describe the mechanism of chemical reactions and enzymatic processes. Thus, it is often used in free energy simulations. The use of such a reaction coordinate also allows convenient analysis to compare specific structures with those obtained from spectroscopic and X-ray diffraction experiments [215]. However, one may sometimes need a more complicated reaction coordinate to yield a more accurate description of the reaction. As one example, the PMF along two reaction coordinates can be obtained by an extension of the one-dimensional Eq. 5.1 to two dimensions [216].

Choice of Reaction Coordinates. For the PMF study in this chapter, a geometrical variable was chosen as the reaction coordinate, r. The distance (r_1) between two atoms, C_α of Trp^{98L} (blue) and the ring oxygen of Abe (the ligand is in orange) was used to study the dissociation of protein-trisaccharide complexes (see Figure 5.1). The positions of these two atoms are shown as yellow and red spheres, respectively.

The reaction coordinate was chosen based on the scFv-trisaccharide ligand interaction map in the gas phase (see Figure 5.2). The pair of Trp^{98L} and Abe indicates a specific, strong interaction with average bond length $r = 2.98$ Å, angle $\theta = 157.69°$

Figure 5.1: The reaction coordinate used in free energy calculations.

and 99 % occupancy from structural H-bond analysis [162].

5.2.4 Restraints and convergence

The central question in free energy simulations is whether convergence is achieved. That is, simulations must have enough time to sample all of the relevant regions of configuration space so that thermodynamic averages will be accurate (converged). However, simulations do not even visit other relevant regions of configuration space if simulations are short and the system is trapped in a metastable state.

In the past decade, restraint potentials have been introduced and employed to avoid convergence problems in free energy simulations [197, 217, 218, 219, 220]. Convergence can be affected by the choice of restraints. Without restraints, the ligand must sample all degrees of freedom and the entire simulation volume during the process of dissociating the ligand from the protein.

The simplest restraint is a single distance restraint between the protein and ligand. When only the protein-ligand distance is restrained the ligand must at least remain

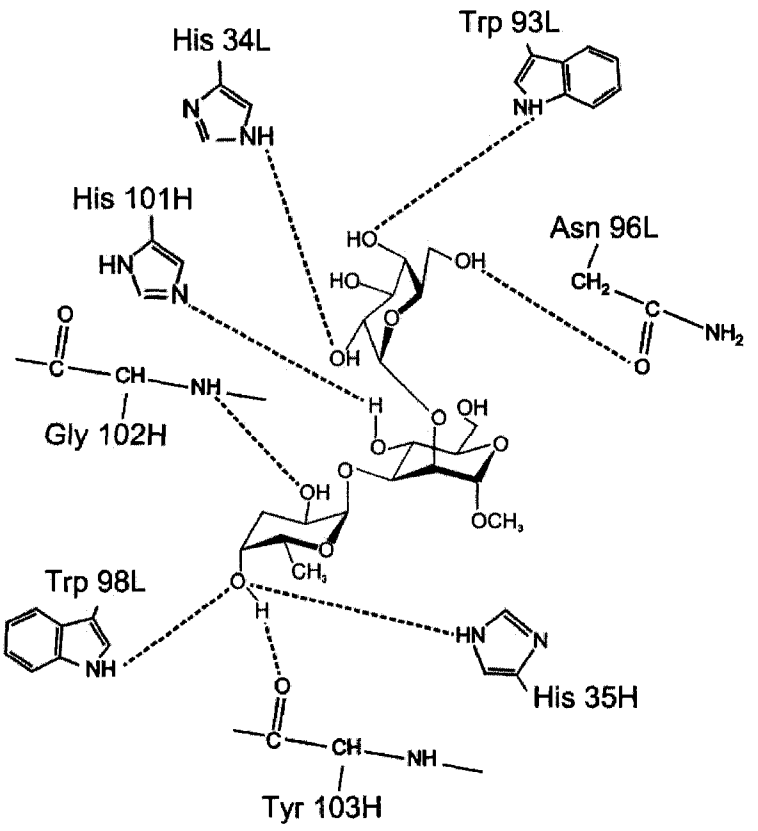


Figure 5.2: Gas phase interaction map determined from MD simulations performed on the (scFv + **1**)$^{8+}$ [162].

near the binding site, but all relevant orientations must still be sampled in every simulation. To improve convergence, orientational restraints have also been used to restrict the ligand's orientation relative to the protein [197, 183, 217]. These restraint potentials further contribute to reduce the amount of configurational space that needs to be thoroughly sampled.

Choice of Restraints. The coordinate system for specifying the overall relative position and orientation of the ligand with respect to the protein was constructed by choosing three atoms within the protein (green) and in the ligand (red) (see Figure 5.3). The three atoms, P_1, P_2 and P_3, labelled in blue, for the scFv of a monoclonal antibody were given by the C_α of Trp^{93L}, Gly^{102H} and Trp^{98L}, respectively. The three

atoms, L_1, L_2 and L_3, for the trisaccharide ligand were given by the ring oxygen of Gal, Man and Abe, respectively. These positions are shown as yellow spheres.

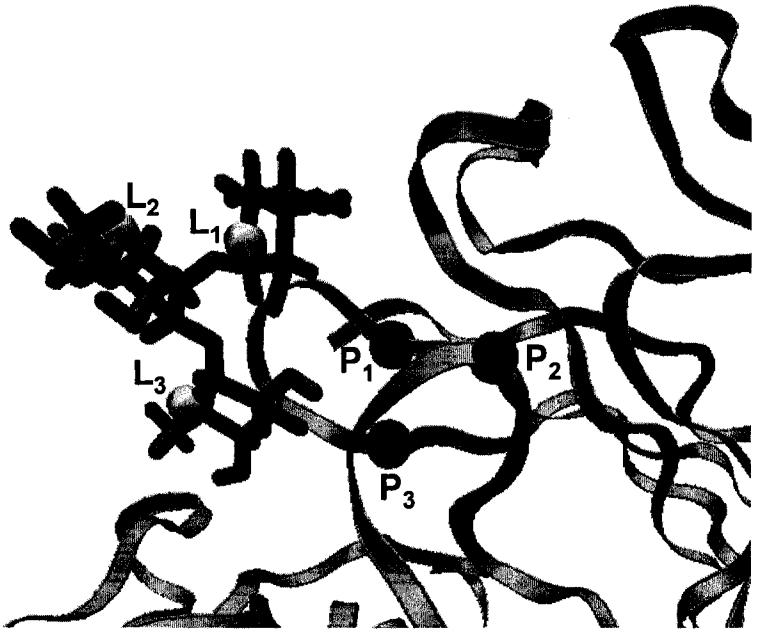


Figure 5.3: The coordinate system used to define the positional and orientational restraints on the ligand.

Six potentials restraining the position and orientation of the ligand were employed in the form of harmonic biasing potentials to help enhance the convergence of the calculations by biasing the ligand to be near its bound configuration as it becomes completely dissociated from its complex. The translational restraint potential is defined as

$$u_t = k_r(r_1 - r_0)^2 + k_a(\theta - \theta_0)^2 + k_a(\phi - \phi_0)^2 \ , \tag{5.5}$$

where r_1 is the distance P_3 - L_3, θ is the angle P_2-P_1-L_1, and ϕ is the dihedral angle P_3-P_2-P_1-L_1. The k_r and k_a are the force constants, and r_0, θ_0 and ϕ_0 are the average values for the bound ligand taken as the reference. Similarly, the orientation of the

ligand was restrained by using the following potential:

$$u_o = k_o(\Theta - \Theta_0)^2 + k_o(\Phi - \Phi_0)^2 + k_o(\Psi - \Psi_0)^2 \ , \tag{5.6}$$

where Θ is angle P_2-L_1-L_2, Φ is the dihedral angle P_2-P_1-L_1-L_2 and Ψ is the dihedral angle P_1-L_1-L_2-L_3. The k_o is the force constant, and Θ_0, Φ_0 and Ψ_0 correspond to the average orientation of the bound ligand. In general, the reference values and the force constants for each restraint potential are determined from the average values based on an unbiased simulation. The magnitude of the force constants is obtained from the fluctuations of its associated coordinates as [217]

$$k_x \approx \frac{k_B T}{\langle \Delta x^2 \rangle} \ . \tag{5.7}$$

The units of force constants used in distance and angle restraint potentials are kcal/mol Å^2 and kcal/mol per rad^2, respectively.

5.3 Methods

All MD simulations were carried out with the recently released AMBER 10 program suite [221]. The crystal structure (1MFA) [13] was used for the initial geometry of the (scFv + **1**) complex. The simulations were performed using the AMBER 94 forcefield with the GLYCAM parameter set for oligosaccharides [12]. Electrostatic potential (ESP) atomic partial charges, determined by Woods and co-workers [60], were used for **1**. The (scFv + **1**) complex at the +8 charge state was chosen for investigation. A series of mutants were also selected: a single amino acid modification of the protein (active site mutation: His^{101H}Ala); a single modification of the ligand (functional group modification: αGal[αAbe](4-deoxyαMan) (**2**)); and simultaneous modification of the protein and the ligand (dual modification: His^{101H}Ala-αGal[αAbe](4-deoxyαMan)). Figure 5.4 shows the structures of the native trisaccharide ligand (**1**) and its monodeoxy analog (**2**).

To get the reference distance, angle values, and force constants for the biasing potentials, unrestrained simulations of the fully interacting ligand in the binding site were performed for the unmodified complex and its mutants. The energies of the

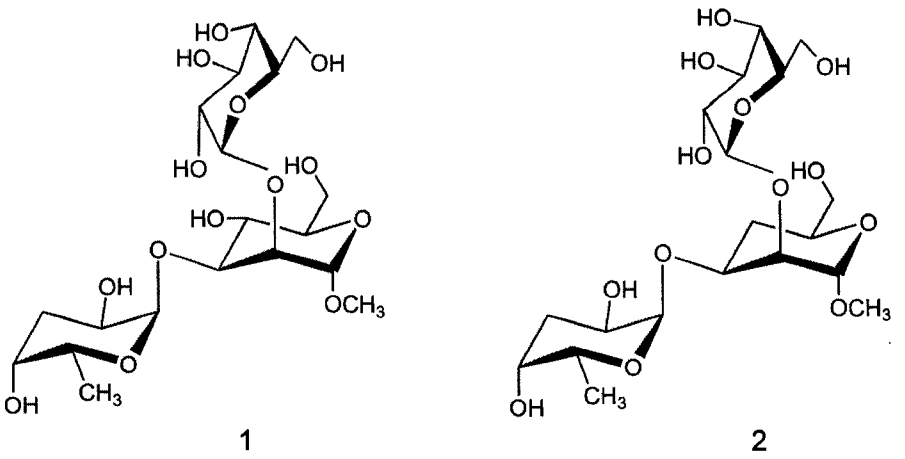


Figure 5.4: Structures of the trisaccharide ligands (**1** and **2**)

(scFv + **1**)$^{8+}$ ion and its mutants were first minimized with the conjugate gradient method using a 0.0001 kcal/mol Å convergence criterion. The entire system was then heated from 10 to 300 K over a period of 15 ps. In order to mimic experimental conditions, simulations were performed in the gas phase under isothermal conditions. Constant temperature was maintained using the weak-coupling algorithm with time constant of 1.0 ps [152]. During the simulation, bond length constraints were applied to all hydrogen-containing bonds using the SHAKE algorithm [24]. The system was equilibrated for 1 ns with a time step of 1 fs. After this period, production dynamics were performed for 4 ns and data were collected every 500 fs.

The PMF along the distance (r_1) was calculated by using umbrella sampling simulations [192, 41] and the WHAM method for unbiasing the data from multiple simulations [42, 193, 41]. After a 1 ns period of the above equilibration, initial configurations for umbrella sampling were then generated in the presence of the restraint potentials. The PMF was calculated with a series of simulations in different windows; 49 windows centered at 0.5 Å intervals from r_1 = 6.0 - 30.0 Å. For each window, the system was further equilibrated for 10 ps, followed by production for 90 ps with a time step of 1 fs. The Langevin algorithm was employed to maintain constant temperature with the collision frequency $\gamma = 5$ ps^{-1} [222].

To provide correct PMF results, the effect of the restraint potentials was unbiased using the WHAM approach. It should be noted that the data were unbiased only for the distance restraint potential, but not for angle or dihedral angle restraint potentials. Since we are interested in the relative difference in rate constants between the unmodified and the mutant complexes, the resulting biased data do not affect the ultimate quantities of interest.

5.4 Results

In addition to the convergence of simulations mentioned in Section 5.2.4, equilibration is another important condition to be met. The system evolves from the starting configuration to reach equilibrium, and equilibrium should continue until the values of a set of monitored properties become stable [22]. Usually, the energy, temperature, pressure and the structural properties are used to monitor the progress of the equilibration. Thus, before we carry the analysis of the trajectories, we first assessed whether the system has reached equilibrium. Here, for each window, the r_1 along the simulation time is used to assess equilibration. An illustration is given in Figure 5.5. Small fluctuations of r_1 about the restrained distance for each window indicate that the system reaches equilibrium after $\sim$2 ps and that the 10 ps of equilibration time used in our simulations is therefore sufficient.

Next, the umbrella sampling technique requires that adjacent windows exhibit some overlap in the distributions of r_1 in order to obtain the potential of mean force [41]. The distributions of some r_1 values (the first 9 windows at r_1 = 6.0 - 10 Å) are shown in Figure 5.6 to illustrate this overlap.

By unbiasing and recombining the results of all the different simulations (windows) using WHAM [42], the final PMF estimate, $w(r_1)$, was obtained. It is worth noting that the data were unbiased only for the distance restraint potential as stated in Section 5.3. The calculated PMFs for the unmodified complex and its mutants are plotted in Figure 5.7. We observed that the activation barrier is lower for the mutants because mutations exclude the possibility of the interactions between the protein and

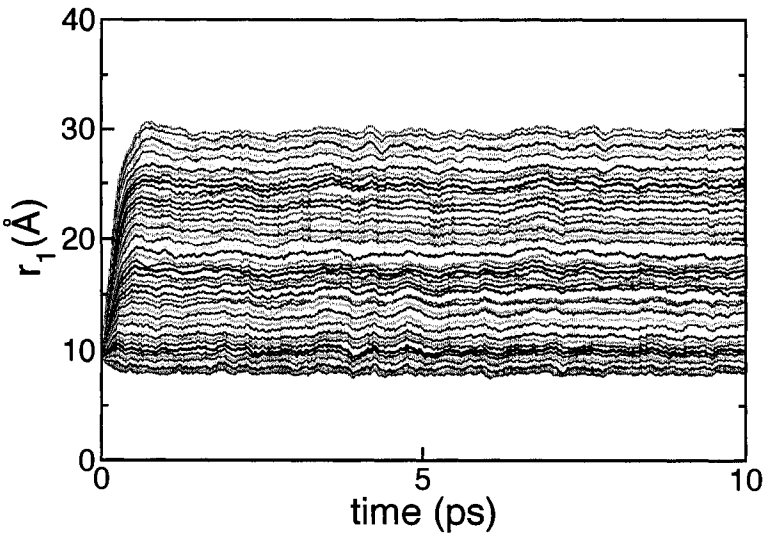


Figure 5.5: Equilibration trajectories for the αGal[αAbe](4-deoxyαMan) mutant (each colour corresponds to a different window).

ligand. This effect is due to the replacement of interacting functional groups on protein, or sugar, or both, by non-interacting substituents, thus lowering the affinity of the ligand for a protein.

Table 5.1: The calculated and experimental dissociation rate constants for the unmodified complex and its mutants at $T = 300$ K.

complex	$k_{r\dagger}^{TST}$ (s^{-1})	k_{exp} (s^{-1})
unmodified complex	3.3×10^{-23}	$4.6 \times 10^{-13} \pm 14.4 \times 10^{-13}$
$His^{101H}Ala$	1.0×10^{-17}	$2.1 \times 10^{-12} \pm 4.3 \times 10^{-12}$
αGal[αAbe](4-deoxyαMan)	8.7×10^{-23}	$6.2 \times 10^{-12} \pm 11.8 \times 10^{-12}$
His^{101H}Ala-αGal[αAbe](4-deoxyαMan)	8.6×10^{-16}	$5.7 \times 10^{-12} \pm 6.9 \times 10^{-12}$

The calculated ($k_{r\dagger}^{TST}$) and experimental (k_{exp}) dissociation rate constants for the unmodified complex and its mutants at $T = 300$ K are provided in Table 5.1. Note that no minimum was found for the $-\left(\frac{dA(r)}{dr}\right)_{r\dagger}$ function discussed in Section 5.2.2 in all the cases under study. A common dividing surface radius $r^{\dagger} = 25$ Å was therefore chosen for all cases. This value is near the top of the barrier and corresponds to the

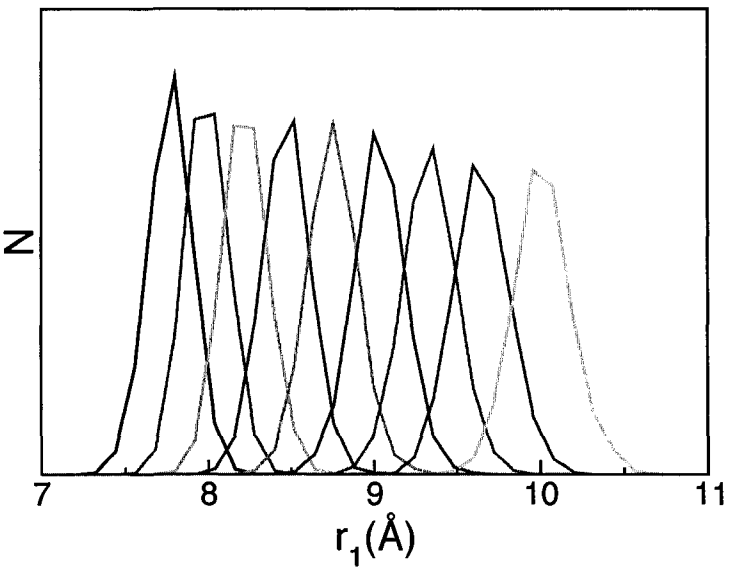


Figure 5.6: Distributions of r_1 (the first 9 windows separated by 0.5 Å for r_1 = 6.0 - 10 Å , N = Number of occurrences)

dissociated complex. The dissociation of the protein-ligand complex occurs faster for the mutants than the unmodified complex as expected.

This big discrepancy between our preliminary computational results and experimental rate constants could be attributed to the length of simulation and the choice of $r^{\dagger}$. Longer simulation time, convergence studies and careful choice of $r^{\dagger}$ are strongly encouraged for future investigations.

5.5 Conclusions

In summary, we have computed the potential of mean forces for the unmodified protein-ligand complex and its mutants in the gas phase at T = 300 K. Restraint potentials were employed in the calculation of the protein-ligand binding free energy to avoid convergence problems. Our study showed that the potential of mean force barrier is lowered (smaller free energy of activation) in the case of mutants. The dissociation rate constants were also calculated from the potentials of mean force by employing TST. The dissociation happens on a faster timescale for the mutants

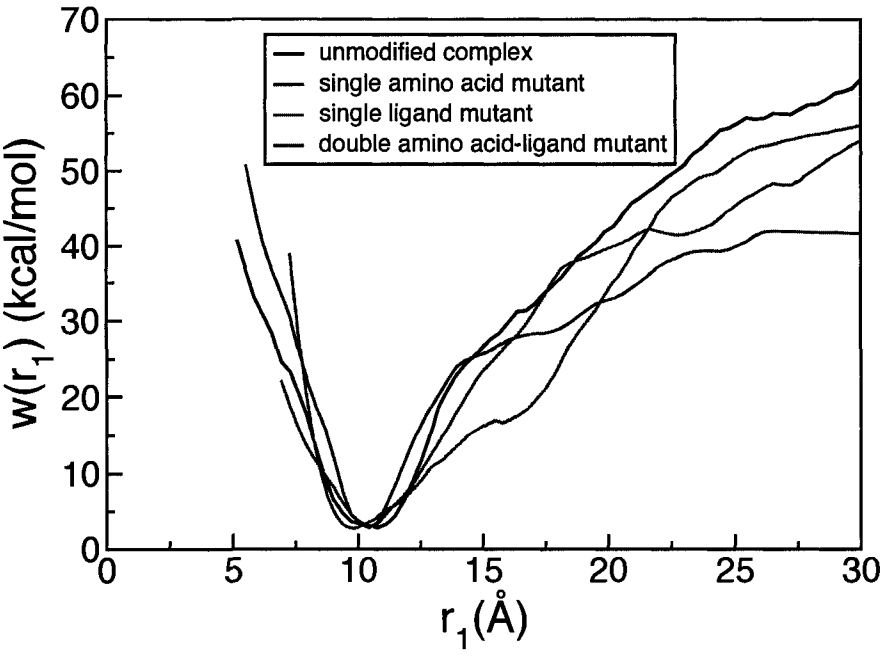


Figure 5.7: Potential of mean force $w(r_1)$ plots of the unmodified complex and mutants along a reaction coordinate r_1 at $T = 300$ K: the unmodified complex, black line; His^{101H}Ala mutant, red line; αGal[αAbe](4-deoxyαMan) mutant, green line; and His^{101H}Ala-αGal[αAbe](4-deoxyαMan) mutant, blue line.

and this can be explained by the removal of interactions between the protein and ligand through mutations. Further studies on computing potentials of mean force and calculating the rate constants for the unmodified complex and its mutants at different temperatures (400 and 500 K) are underway. The resulting dissociation rate constants will then be fitted to the Arrhenius relation in order to obtain pre-exponential factors and activation energies. The dissociation kinetics and energetics computed theoretically will then be compared directly with available experimental data [162].

Chapter 6

Conclusions

Molecular recognition plays a crucial role in a wide range of biological processes. The recognition of carbohydrates by proteins is a subject of major interest with many practical implications once the specifics of the molecular recognition process are understood. In order to understand carbohydrate recognition more fully, a deeper appreciation of intermolecular interactions that govern the affinity and specificity of the carbohydrate-protein binding is required. Over many years, a great deal of information about carbohydrate-protein binding has been obtained from both experimental and theoretical methods. The contributions of many researchers, including ours, have helped to draw a more detailed picture of the recognition process.

In Chapter 1 we briefly reviewed experimental studies on protein-carbohydrate interactions. We also pointed out some of difficulties present in these experimental methods and offered computational approaches as a promising alternative means of investigating carbohydrate-protein recognition processes at a molecular level. In addition, we provided a description of MD simulations technique used for the studies presented in this thesis. Throughout this work, we focused on studies that aim to achieve a better understanding of carbohydrate-protein recognition at a microscopic level in terms of the structure and dynamics of carbohydrate-protein complexes by means of MD simulations. The investigations undertaken of this thesis can be categorized into two main aspects: (i) the development of tools and simulation protocols to contribute to the modelling of carbohydrates; (ii) the investigation of the factors affecting the nature of the intrinsic binding interactions between a protein and a

carbohydrate molecule. The main findings presented in this thesis are summarized below.

In Chapter 2 we developed a new approach to model furanosides in solution. Due to the inherent flexibility of furanosides, the conformations they can adopt are diverse, and this makes their conformational analysis much more complicated than similar studies conducted on the more rigid pyranosides. We proposed a new charge derivation approach that accounts for the flexibility of these ring systems by taking an average of the charges from a large number of conformers. The first test of this approach was performed on the methyl-α-D-arabinofuranoside. It was shown that the model can predict conformational properties with good agreement with NMR experimental data. From the knowledge of solution conformations of sugars, we proceeded to investigate the interactions between sugars and proteins as discussed below.

We presented in Chapter 3 a detailed study of the intermolecular interactions of a charged protein-ligand complex in the gas phase. A single chain-variable domain fragment (scFv) of a carbohydrate-binding antibody and its native trisaccharide αGal[αAbe]αMan served as a model system. Simulations were carried out for the protonated (+8) and deprotonated (-8) ions of a complex in order to predict which specific interactions are preserved under the conditions of mass spectrometric experiments. Intermolecular H-bonds were identified and gas-phase maps were generated for the complex. This was compared with the corresponding experimentally-derived maps. Most of the intermolecular interactions identified from our simulations of the protonated ion of the complex were also observed experimentally; the agreement was less favourable in the case of deprotonated ion. However, both the simulation and experimental results pointed to structural differences between the +8 and -8 ions. In addition, comparison of these gas-phase H-bond maps with the crystal structure of the complex provided a deeper understanding of the structural changes that accompany the transfer of complexes from solution to the gas phase. First, at least two of the specific H-bonds are conserved upon transfer of the complex from solution to the gas phase. This is compelling evidence for the retention of specific interactions. In addition, newly created (nonspecific) interactions were identified. We also found that

the water-mediated H-bonds identified in the crystal structure are lost upon transfer of the complexes from solution to the gas phase and are replaced with direct H-bonds.

To further our understanding of protein-ligand interactions we focused on the behaviour of "crystallographic" water molecules. In Chapter 4, as opposed to the previous work on equilibrium properties in the form of the identification of H-bonds in Chapter 3, we studied the temperature dependence of the *dynamics* of the three "crystallographic" water molecules present in a protein-ligand complex at +8 charge state. The computed water densities showed that the most buried water molecule remains localized. However, the other two become more and more delocalized as the temperature is increased up to 300 K, and finally they diffuse away from the binding site and interact with the surface of the protein at this temperature. The structural relaxation of H-bonds between water molecules and the complex was then investigated in terms of hydrogen bond lifetime dynamics to better understand the role of water molecules in protein-ligand binding. Generally, the H-bond lifetimes decrease as temperature increases. The structural relaxation of the H-bonds formed at high temperature is faster than that at low temperature. In addition, the observation of the two water molecules diffusing away from the binding site at high temperatures in the water densities is confirmed by their shorter H-bond lifetimes.

Finally, in Chapter 5, we presented the results of the computation of the free energy profile along a protein-ligand complex dissociation coordinate. Potentials of mean force were calculated for the native complex and its mutants. The contribution of a specific protein-ligand interaction to the stability of the mutant complexes were understood in terms of changes in the potential of mean force barrier height. Our calculations showed that the potential of mean force barrier is lowered in the case of mutants. The mutant from double modification showed the lowest potential of mean force barrier. The dissociation rate constants were also obtained, and showed that the dissociation of the complex occurs faster for the mutants. The various possibilities of extending the present work is discussed in the final section of the chapter.

6.1 Contributions to Research Tools and Original Knowledge

In this section, we provide our novel contributions to scientific knowledge and computational methodologies. Our first accomplishment is the development of an approach to derive charges for furanose rings. Several models have been developed and implemented to perform simulation studies on carbohydrates [11, 12]. The most notable feature that distinguishes this new approach from the existing model is the incorporation of the inherent flexibility of the furanoside rings. Following the successful implementation of this approach to model methyl-α-D-arabinofuranoside, this method was and is currently being used in the study of other commonly occurring furanoside monosaccharides (*e.g.* β-D-arabinofuranoside and β-D-galactofuranoside). The extension of this method also involves the study of more complex oligomeric and polymeric structures.

Our investigation of the interactions between proteins and ligands adds to the existing multitude of reports on the exploration of systems comprised of proteins or protein complexes through computational simulations in tandem with experimental means. The specific combination of simulations in the gas-phase and mass spectrometric experiments however, represents a novel tool for probing directly the interactions of protein-ligand systems. In particular, we have developed a simulation protocol to study oligosaccharides and their complexes with proteins in the gas phase. The methodology was applied to determine structural properties such as H-bond distances and angles, dynamical properties of H-bonds, energetic properties in the form of binding free energy, and kinetics of protein-ligand dissociation. The simulation results were compared with experimentally determined quantities to both rationalize experimental observations and act in a predictive capacity. Due to the computational nature of the developed method, it possesses an inherent ability to offer a broader picture than experimental techniques, which are often affected by practical limitations. For instance, the mass spectrometric tool used in investigating the protein-ligand complex relied on mutagenesis, a technique involving the mutation of interacting

sites. In practice, the mutation was sometimes observed to lead to a dissociation of the complex, thus restricting the scope of the experimental investigation. In contrast, our computational approach is unaffected by such problems. We are of course aware that any simulation approach is also limited, especially because of uncertainties in the models.

In addition, a large portion of the work involved in this thesis has required the development of computer codes and analysis software. Computer software codes were designed and programmed,[1] *e.g.* scripts for the calculation of H-bond lifetime correlation functions and the reaction rate constants, and an automated input generators for umbrella sampling.

6.2 Future Directions and Outlook

In this last section of the thesis, we propose some possible future research avenues.

6.2.1 Dissociation kinetics of a protein-ligand complex: Arrhenius analysis

In Chapter 5, we mentioned that theoretically determined dissociation kinetics and energetics could be directly compared to experiment via free energy calculations. Dissociation rate constants (k) for the unmodified complex and its mutants can be obtained from the calculated PMFs. Further studies on computing PMFs and rate constants for the unmodified complex (scFv-αGal[αAbe]αMan) and its mutants at different temperatures (T = 400 and 500 K) are actually in progress. A series of additional mutants could be considered: a single amino acid modification of the protein (active site mutation: His^{101H}Ala, His^{34L}Ala, His^{35H}Ala, His^{97L}Ala, Asn^{96L}Ala); a single modification of the ligand (functional group modification: αGal[αAbe](4-deoxyαMan), αGal[αAbe](6-deoxyαMan)); and simultaneous modification of the protein and the ligand (dual modification: His^{101H}Ala-αGal[αAbe](4-deoxyαMan)). The resulting dissociation rate constants can be fitted to the Arrhenius relation. Arrhe-

[1]For the work in this thesis, over 5000 lines of code were written using *python* language.

nius plots are constructed from the computed temperature-dependent rate constants by plotting $\ln k$ vs $\frac{1}{T}$. This procedure is simply derived from the Arrhenius expression $k = A\exp(\frac{-E_a}{RT})$ where A is the so-called pre-exponential factor and E_a is the activation energy. By rearranging the above equation, a plot of $\ln k$ vs. $\frac{1}{T}$ has slope $\frac{-E_a}{R}$ and intercept $\ln A$ [204]. Thus, the calculated E_a and A will provide a direct comparison with available experimental data obtained from blackbody infrared radiative dissociation (BIRD) technique [162, 128, 151, 200].

6.2.2 Water evaporation

As an extension of the study on the H-bond dynamics of "crystallographic" water molecules, we suggest free energy calculations of water-complex dissociation in the gas phase to calculate water evaporation rate constants. PMFs can be calculated for the dissociation of an individual water molecule from the complex along a chosen reaction coordinate at different temperatures. The distance between the center-of-mass of a water molecule and one of the key amino acid residues or an hydroxyl group on the ligand in the binding site can be selected as a possible reaction coordinate. These PMF calculations would allow us to determine how long water molecules stay in the complex at a given temperature, *i.e.* the time scale of the water evaporation for an individual water molecule, by calculating evaporation rate constants using transition state theory. It would be interesting to compare the timescales of these water evaporation rates with the ones of the H-bond dynamical analysis of Chapter 4.

6.2.3 Dissociation kinetics of a protein-ligand complex in solution

The next class of systems we propose to study is protein-ligand complexes in solution. Potentials of mean force can be calculated for the dissociation of these systems. Comparison of the free energies of the complex in the gas phase with the ones in solution will provide new insights into the solvent effects that accompany the transfer of the complexes from solution to the gas phase. In our research group, the comparison

of the PMF for the dissociation of an hexasaccharide bound to an arabinan binding protein was calculated both in the gas phase and solution. Preliminary results are shown in Figure 6.1. This study shows that the PMF barrier is lowered when one goes from the gas phase to solution due to the stabilizing effects of the solvent [223]. With these kinds of solution simulations, one can also envisage the design of novel ligands or more potent inhibitors.

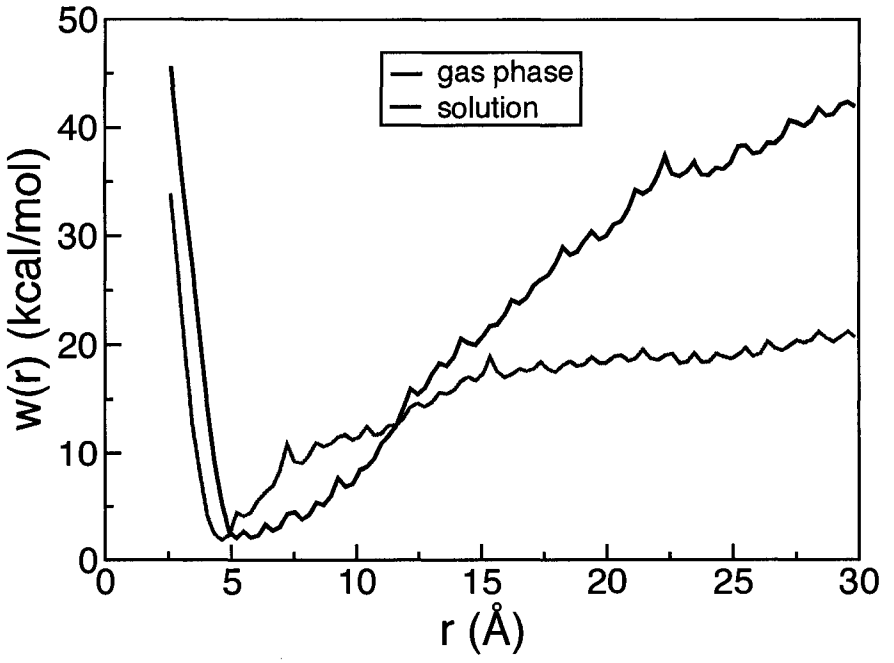


Figure 6.1: Potential of mean force of the protein-hexasaccharide complex along a reaction coordinate r at $T = 300$ K in the gas phase and solution

Bibliography

[1] K. Rippe, http://www.kip.uni-heidelberg.de/chromcon/teaching/index_teaching.html.

[2] R. U. Lemieux, Chem. Soc. Rev. **18**, 347 (1989).

[3] L. Lasky, Science **258**, 964 (1992).

[4] R. A. Dwek, Chem. Rev. **96**, 683 (1996).

[5] M. R. Wormald et al., Chem. Rev. **102**, 371 (2002).

[6] J. L. Asensio et al., Glycobiology **8**, 569 (1998).

[7] G. Colombo, M. Meli, J. Canada, J. L. Asensio, and J. Jimenez-Barbero, Carb. Res. **340**, 1039 (2005).

[8] J. Gonzalez-Outeirino, K. N. Kirschner, S. Thobhani, and R. J. Woods, Can. J. Chem. **84**, 569 (2006).

[9] W. A. Bubb, Concept. Magn. Res. **19A**, 1 (2003).

[10] J. Jiménez-Barbero and T. Peters, *NMR of Glycoconjugates* (Wiley-VCH, Weinheim, 2002).

[11] K. N. Kirschner et al., J. Comput. Chem. **29**, 622 (2008).

[12] R. J. Woods, R. A. Dwek, C. J. Edge, and B. Frase-Reid, J. Phys. Chem. **99**, 3832 (1995).

[13] A. Zdanov et al., Proc. Natl. Acad. Sci. (U.S.A.) **91**, 6423 (1994).

[14] M. Cygler, D. R. Rose, and D. R. Bundle, Science **253**, 442 (1991).

[15] I. N. Levine, *Physical Chemistry* (McGraw-Hill, New York, 2002).

[16] M. Born and R. Oppenheimer, Ann. Phys. (Berlin) **84**, 457 (1927).

[17] H. Goldstein, C. Poole, and J. Safko, *Classical Mechanics* (Addison-Wesley, Sanfrancisco, 2002).

[18] N. L. Doltsinis, Comput. Nanosci., NIC Series **31**, 389 (2006).

[19] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic Press, Inc., 1996).

[20] D. A. McQuarrie, *Statistical Mechanics* (University Science Books, 1973).

[21] L. Verlet, Phys. Rev. **159**, 98 (1967).

[22] A. R. Leach, *Molecular Modelling, Principles and Applications* (Pearson Education Limited, 2001).

[23] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, 1987).

[24] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berndsen, J. Comput. Phys **23**, 327 (1977).

[25] H. J. C. Berendsen, Comp. Phys. Commun. **44**, 233 (1987).

[26] R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 5471 (1985).

[27] S. Nosé, J. Phys.: Condens. Matter **2**, SA115 (1990).

[28] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).

[29] S. A. Adelman and J. D. Doll, J. Chem. Phys. **64**, 2375 (1976).

[30] H. C. Andersen, J. Chem. Phys. **72**, 2384 (1980).

[31] S. Nosé, Mol. Phys. , 255 (1984).

[32] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).

[33] W. G. Hoover, Phys. Rev. A **34**, 2499 (1986).

[34] D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford University Press, 1987).

[35] L. Onsager, Phys. Rev. **37**, 405 (1931).

[36] R. Zwanzig, Annu. Rev. Phys. Chem. **16**, 67 (1965).

[37] P. Kollman, Chem. Rev. **93**, 2395 (1993).

[38] W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson, *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications* (Escom, Leiden, 1993).

[39] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).

[40] L. G., X. Zhang, and Q. Cui, J. Phys. Chem. B **107**, 8643 (2003).

[41] B. Roux, Comput. Phys. Comm. **91**, 275 (1995).

[42] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, J. Comput. Chem. **13**, 1011 (1992).

[43] A. M. Ferrenberg, Phys. Rev. Lett. **63**, 1195 (1989).

[44] S. Strogatz, *In What is Your Dangerous Idea? (J. Brockman, ed)* (Harper-Collins, 2006).

[45] N. L. Allinger, J. Am. Chem. Soc. **99**, 3279 (1977).

[46] N. L. Allinger, Y. H. Yuh, and J.-H. Lii, J. Am. Chem. Soc. **111**, 8551 (1989).

[47] W. D. Cornell et al., J. Am. Chem. Soc. **117**, 5179 (1995).

[48] B. R. Brooks et al., J. Comput. Chem. **4**, 184 (1983).

[49] W. Damm, J. Tirado-Rives, and W. L. Jorgensen, J. Am. Chem. Soc. **118**, 11225 (1996).

[50] L. D. Schuler, X. Daura, and W. F. van Gunsteren, J. Comp. Chem. **22**, 1205 (2001).

[51] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman, J. Phys. Chem. **97**, 10269 (1993).

[52] T. Peters and B. M. Pinto, Curr. Opin. Struct. Biol. **6**, 710 (1996).

[53] H. Thogersen, R. U. Lemieux, K. Bock, and B. Meyer, Can. J. Chem. **60**, 44 (1982).

[54] N. L. Allinger, J. Am. Chem. Soc. **99**, 8127 (1977).

[55] J.-H. Lii and N. L. Allinger, J. Am. Chem. Soc. **111**, 8566 (1989).

[56] S. J. Weiner et al., J. Am. Chem. Soc. **106**, 765 (1984).

[57] R. J. Woods, Glycocon. J. **15**, 209 (1998).

[58] R. J. Woods, C. J. Edge, and R. A. Dwek, Nature Struct. Biol. **1**, 499 (1994).

[59] R. J. Woods and R. J. Chappelle, J. Mol. Struct. (THEOCHEM) **527**, 149 (2000).

[60] A. Pathiaseril and R. J. Woods, J. Am. Chem. Soc **122**, 331 (2000).

[61] Z. J. Witczak, Curr. Med. Chem. **2**, 392 (1995).

[62] K. K.-S. Ng et al., J. Biol. Chem. **288**, 16088 (2002).

[63] D. K. Mandal, N. Kishore, and C. F. Brewer, Biochem. **33**, 1149 (1994).

[64] F. P. Schwarz, K. D. Puri, R. G. Bhat, and A. Surolia, J. Biol. Chem. **268**, 7668 (1993).

[65] R. J. Woods, *In Reviews in Computational Chemistry (K. B. Lipkowitz and D. B. Boyd, eds)* (New York: VCH Publishers Inc., 1996).

[66] R. J. Woods, Curr. Opin. Struct. Biol. **5**, 591 (1995).

[67] A. French and V. Tran, Biopolymers **29**, 1599 (1990).

[68] K. N. Kirschner and R. J. Woods, Proc. Natl. Acad. Sci. (U.S.A.) **98**, 10541 (2001).

[69] S. W. Homans, Glycobiology **3**, 551 (1993).

[70] K.-H. Ott and B. Meyer, Carbohydr. Res. **281**, 11 (1996).

[71] K. Ueda and J. W. Brady, Biopolymers **41**, 323 (1997).

[72] N. K. de Vries and H. M. Buck, Carbohydr. Res. **165**, 1 (1987).

[73] P.-E. Jansson, L. Kenne, and I. Kolare, Carbohydr. Res. **257**, 163 (1994).

[74] S. E. Barrows, J. W. Storer, C. J. Cramer, A. D. French, and D. G. Truhlar, J. Comput. Chem. **19**, 1111 (1998).

[75] E. W. Wooten, C. J. Edge, R. Bazzo, R. A. Dwek, and T. W. Rademacher, Carbohydr. Res. **203**, 13 (1990).

[76] J. F. G. Vliegenthart and R. J. Woods, *NMR Spectroscopy and Computer Modeling of Carbohydrates: Recent Advances* (American Chemical Society, Washington, DC, 2006).

[77] D. A. Case et al., J. Comput. Chem. **26**, 1668 (2005).

[78] W. Saenger, *Principles of Nucleic Acid Structure* (Spring-Verlag; Berlin, 1988).

[79] J. B. Houseknecht and T. L. Lowary, Curr. Opin. Struct. Biol. **5**, 677 (2001).

[80] T. L. Lowary, Curr. Opin. Struct. Biol. **7**, 749 (2003).

[81] M. H. Ryder, M. E. Tate, and G. P. Jones, J. Biol. Chem. **165**, 327 (1984).

[82] K. Komatsu, H. Shigemori, and J. Kobayashi, J. Org. Chem. **66**, 6189 (2001).

[83] P. J. Brennan and H. Nikaido, Annu. Rev. Biochem. **64**, 29 (1995).

[84] D. C. Crick, S. Mahaptra, and P. Brennan, Glycobiology **11**, 107R (2001).

[85] V. Briken, S. A. Porcelli, G. S. Besra, and L. Kremer, Mol. Microbiol. **53**, 391 (2004).

[86] J. Jimenez-Barbero, J. L. Asensio, F. J. Canada, and A. Poveda, Curr. Opin. Struct. Biol. **9**, 549 (1999).

[87] C. Altona and M. Sundaralingam, J. Am. Chem. Soc. **94**, 8205 (1972).

[88] C. Altona and M. Sundaralingam, J. Am. Chem. Soc. **95**, 2333 (1973).

[89] J. B. Houseknecht, C. A. C. M. Hadad, and T. L. Lowary, J. Org. Chem. **67**, 4647 (2002).

[90] F. Deleeuw and C. Altona, J. Comput. Chem. **4**, 428 (1983).

[91] R. U. Lemieux, Tetrahedron **30**, 1933 (1974).

[92] S. Wolfe, Acc. Chem. Res. **5**, 102 (1972).

[93] N. K. Devrise and H. M. Buck, Carbohydr. Res. **165**, 1 (1987).

[94] K. Bock and J. O. Duus, J. Carbohydr. Chem. **13**, 513 (1994).

[95] I. Tbaroska and J. P. Carver, J. Phys. Chem. B **101**, 2992 (1997).

[96] F. W. D'Souza, J. Ayers, P. R. McCarren, and T. L. Lowary, J. Am. Chem. Soc. **122**, 1251 (2000).

[97] P. R. McCarren, M. T. Gordon, T. L. Lowary, and C. M. Hadad, J. Phys. Chem. A **105**, 5911 (2001).

[98] M. R. Gordon, T. L. Lowary, and C. M. Hadad, J. Am. Chem. Soc. **121**, 9682 (1999).

[99] J. B. Houseknecht, T. L. Lowary, and C. M. Hadad, J. Phys. Chem. A **107**, 5763 (2003).

[100] M. T. Gordon, T. L. Lowary, and C. M. Hadad, J. Org. Chem. **65**, 4954 (2000).

[101] J. B. Houseknecht, P. R. McCarren, T. L. Lowary, and C. M. Hadad, J. Am. Chem. Soc. **123**, 8811 (2001).

[102] J. B. Houseknecht, T. L. Lowary, and C. M. Hadad, J. Phys. Chem. A **107**, 372 (2003).

[103] S. Cros, C. H. Dupenhoat, S. Perez, and A. Imberty, Carbohydr. Res. **248**, 81 (1993).

[104] S. Cros, A. Imberty, N. Bouchemal, C. H. Dupenhoat, and S. Perez, Biopoymers **34**, 1433 (1994).

[105] M. K. Dowd, A. D. French, and P. J. Reilly, J. Carbohydr. Chem. **19**, 1091 (2000).

[106] K. Mazeau and S. Perez, Carbohydr. Res. **311**, 203 (1998).

[107] A. D. French and M. K. Dowd, J. Comput. Chem. **15**, 561 (1994).

[108] A. D. French, M. K. Dowd, and P. J. Reilly, J. Mol. Struct. THEOCHEM **395**, 271 (1997).

[109] D. A. Case et al., Amber 9, 2006.

[110] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).

[111] M. Basma, S. Sundara, D. Calgan, T. Vernali, and R. J. Woods, J. Comput. Chem. **22**, 1125 (2001).

[112] A. G. Evdokimov, A. J. Kalb, T. F. Koetzle, and W. T. Klooster, J. Phys. Chem. A **103**, 744 (1999).

[113] M. J. Frisch et al., Gaussian, inc., 2004.

[114] T. M. Su and Y. S. Yang, Biochemistry **42**, 6863 (2003).

[115] E. V. Pletneva, A. T. Laederach, D. B. Fulton, and N. M. Kostic, J. Am. Chem. Soc. **26**, 6232 (2001).

[116] N. R. Silvaggi, K. Kaur, S. A. Adediran, R. F. Pratt, and J. A. Kelly, Biochemistry , 7046 (2004).

[117] J. A. Loo, Mass Spectrom. Rev. **16**, 1 (1997).

[118] J. L. P. Benesch, F. Sobott, and C. V. Robinson, Anal. Chem. **75**, 2208 (2003).

[119] K. A. Newton, R. Amunugama, and S. A. McLuckey, J. Phys. Chem. **109**, 3608 (2005).

[120] F. W. McLafferty, Z. Q. Guan, U. Haupts, T. D. Wood, and N. L. Kelleher, J. Am. Chem. Soc. **120**, 4732 (1998).

[121] K. B. Shelimov, D. E. Clemmer, R. R. Hudgins, and M. F. Jarrold, J. Am. Chem. Soc. **119**, 2240 (1997).

[122] J. Oomens et al., Phys. Chem. Chem. Phys. **7**, 1345 (2005).

[123] A. T. Iavarone and J. H. Parks, J. Am. Soc. Chem. **127**, 8606 (2005).

[124] Y. Xie, J. Zhang, S. Yin, and J. A. Loo, J. Am. Chem. Soc. **128**, 14432 (2003).

[125] M. Tešić, J. Wicki, D. K. Y. Poon, S. G. Withers, and D. J. Douglas, J. Am. Soc. Mass Spectrom. **18**, 64 (2007).

[126] R. C. Dunbar and T. B. McMahon, Science **279**, 194 (1998).

[127] W. D. Price, P. D. Schnier, R. A. Jockush, E. F. Strittmatter, and E. R. Williams, J. Am. Chem. Soc. **118**, 10640 (1996).

[128] E. N. Kitova, D. R. Bundle, and J. S. Klassen, J. Am. Chem. Soc. **124** (2002).

[129] E. N. Kitova, D. R. Bundle, and J. S. Klassen, Angew. Chem. Int. Ed. **43**, 4183 (2004).

[130] G. A. Arteca, C. T. Reimann, and O. Tapia, Mass Spectrom. Rev. **20**, 402 (2001).

[131] G. A. Arteca and O. Tapia, Mass Spectrom. Rev. **20**, 402 (2001).

[132] D. T. Kaleta and M. F. Jarrold, J. Phys. Chem. B **107**, 14529 (2003).

[133] D. T. Kaleta and M. F. Jarrold, J. Phys. Chem. A **106**, 9655 (2002).

[134] A. Patriksson, E. Marklund, and D. van der Spoel, Biochemistry **46**, 933 (2007).

[135] R. W. McLafferty, Science **214**, 280 (1981).

[136] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse, Science **426**, 64 (1989).

[137] M. Yamashita and J. B. Fenn, J. Phys. Chem. **88**, 4671 (1984).

[138] M. S. Wilm and M. Mann, Int. J. Mass Spectrom. Ion Processes **136**, 167 (1994).

[139] M. S. Wilm and M. Mann, Anal. Chem. **68**, 1 (1996).

[140] T. W. Hutchens, M. H. Allen, C. M. Li, and T. T. Yip, Febs Letters **309**, 170 (1992).

[141] P. F. Hu, Q. Z. Ye, and J. A. Loo, Anal. Chem. **66**, 4190 (1994).

[142] X. Cheng et al., J. Am. Chem. Soc. **117**, 8859 (1995).

[143] J. A. Loo et al., Proteins: Struct. Funct. Genet. Suppl. **2**, 28 (1998).

[144] S. M. Blair, E. C. Kempen, and J. S. Brodbelt, J. Am. Soc. Mass Spectrom. **9**, 1049 (1998).

[145] M. J. Greig, H. Gaus, L. L. Cummins, H. Sasmor, and R. H. Griffey, J. Am. Chem. Soc. **117**, 10765 (1995).

[146] J. A. Loo et al., J. Am. Soc. Mass Spectrom. **8**, 234 (1997).

[147] C. L. Hunter, A. G. Mauk, and D. J. Douglas, Biochemistry **36**, 1018 (1997).

[148] P. D. Shinier, J. S. Klassen, E. F. Strittmatter, and E. R. Williams, J. Am. Chem. Soc. **120**, 9605 (1998).

[149] Q. Wu et al., J. Am. Chem. Soc. **119**, 1157 (1997).

[150] J. M. Daniel, S. D. Firess, S. Rajagopalan, S. Wendt, and R. Zenobi, Int. J. Mass Spectrom. **216**, 1 (2002).

[151] E. N. Kitova, D. R. Bundle, and J. S. Klassen, J. Am. Chem. Soc. **124**, 9340 (2002).

[152] H. J. C. Berndsen, J. P. Postma, W. F. Vangunsteren, A. Dinola, and J. R. Haak, J. Chem. Phys **81**, 3684 (1984).

[153] E. P. Hunter and S. G. Lias, Phys. Chem. Ref. Data **37**, 413 (1998).

[154] M. T. Bowers, *Gas Phase Ion Chemistry* (Academic Press, New York, 1979).

[155] D. Bundle et al., Biochemistry. **33**, 5183 (1994).

[156] D. W. Cleveland, Y. Mao, and K. F. Sullivan, Cell **112**, 407 (2003).

[157] A. Helnius and M. Aebi, Science **23**, 2364 (2001).

[158] T. Igakura et al., Science **299**, 1713 (2003).

[159] H.-J. Böhm and G. Schneider, *Protein-Ligand Interactions From Molecular Recoginition to Drug Design* (Wiley-VCH, Weinheim, 2003).

[160] M. G. McCammon et al., Structure **10**, 851 (2002).

[161] J. A. Aquilina, J. L. P. Benesch, O. W. Bateman, C. Slingsby, and C. V. Robinson, Proc. Natl. Acad. Sci. U.S.A. **100**, 10611 (2003).

[162] E. N. Kitova, M. Seo, P.-N. Roy, and J. S. Klassen, J. Am. Chem. Soc. **130**, 1214 (2008).

[163] E. N. Kitova, W. Wang, D. R. Bundle, and J. S. Klassen, J. Am. Chem. Soc. **124**, 13980 (2002).

[164] U. Kaatze, H. Gerke, and R. Pottel, J. Phys. Chem. **90**, 5464 (1986).

[165] A. Shimizu, K. Fumino, K. Yukiyasu, and Y. Taniguchi, J. Mol. Liq. **85**, 269 (2000).

[166] D. I. Svergun et al., Proc. Natl. Acad. Sci. (U.S.A.) **95**, 2267 (1998).

[167] U. Heugen et al., Proc. Natl. Acad. Sci. (U.S.A.) **103**, 12301 (2006).

[168] R. M. Raschke, Curr. Opin. Struct. Biol. **16**, 152 (2006).

[169] R. H. Henchman and J. A. McCammon, Prot. Sci. **11**, 2080 (2002).

[170] A. Luzar, J. Chem. Phys. **113**, 10663 (2000).

[171] A. Luzar and D. Chandler, Nature **379**, 55 (1996).

[172] A. Luzar and D. Chandler, Phys. Rev. Lett. **76**, 928 (1996).

[173] S. Pal, S. Balasubramanian, and B. Bagchi, J. Chem. Phys **117**, 2852 (2002).

[174] S. L. Lee and P. G. Debenedetti, J. Chem. Phys. **122**, 204511 (2005).

[175] M. T. C. M. Costa, Carb. Res. **340**, 2185 (2005).

[176] S. Balasubramanian and B. Bagchi, J. Phys. Chem. B **105**, 12529 (2001).

[177] S. Balasubramanian and B. Bagchi, J. Phys. Chem. B **106**, 3668 (2002).

[178] S. Bandyopadhyay, S. Chakraborty, S. Balasubramanian, and B. Bagchi, J. Am. Chem. Soc. **127**, 4071 (2005).

[179] S. Bandyopadhyay, S. Chakraborty, and B. Bagchi, J. Am. Chem. Soc. , 16660 (2005).

[180] S. Pal, P. Maiti, and B. Bagchi, J. Chem. Phys **125**, 234903 (2006).

[181] Y. Li, H. Li, F. Yang, and S. J. Smith-Gill, Nat. Struct. Biol. **10**, 482 (2003).

[182] G. J. Kroon, H. Mo, M. A. Martinez-Yamount, H. J. Dyson, and P. E. Wright, Protein Sci. **12**, 1386 (2003).

[183] H.-J. Woo and B. Roux, Proc. Natl. Acad. Sci. (U.S.A.) **102**, 6825.

[184] I. F. Thorpe and C. L. B. III, Proc. Natl. Acad. Sci. (U.S.A.) **104**, 8821 (2007).

[185] Y. Deng and B. Roux, J. Chem. Theory Comput. **2**, 1255 (2006).

[186] H. Luo and K. Sharp, Proc. Natl. Acad. Sci. (U.S.A.) **99**, 10399 (2002).

[187] M. K. Gilson and H.-X. Zhou, Annu. Rev. Biophys. Biomol. Struct. **36**, 21 (2007).

[188] D. Trzesniak, A.-P. E. Kunz, and W. F. van Gunsteren, Chem. Phys. Chem. **8**, 162 (2007).

[189] W. L. Jorgensen and C. Ravimohan, J. Chem. Phys. **83**, 3050 (1985).

[190] J. K. Buckner and W. L. Jorgensen, J. Am. Chem. Soc. **111**, 2507 (1989).

[191] D. A. Pearlman, J. Chem. Phys. **98**, 8946 (1993).

[192] J. P. V. G. M. Torrie, J. Comput. Phys. **23**, 187 (1977).

[193] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, J. Comput. Chem. **16**, 1339 (1995).

[194] J. W. Essex, D. L. Severance, J. Tirado-Rives, and W. L. Jorgensen, J. Phys. Chem. B **101**, 9663 (1997).

[195] S. Miyamoto and P. A. Kollman, Proteins: Struct. Funct. Bioinf. **16**, 226 (1993).

[196] T. Fox, T. S. Scanlan, and P. A. Kollman, J. Am. Chem. Soc. **119**, 11571 (1997).

[197] D. L. Mobley, J. D. Chodera, and K. A. Dill, J. Chem. Phys. **125**, 084902 (2006).

[198] D. R. Bundle et al., Biochem. **33**, 5172 (1994).

[199] W. D. Price, P. D. Schnier, and E. R. Williams, Anal. Chem. **68**, 859 (1996).

[200] J. S. Klassen, P. D. Schnier, and E. R. Williams, J. Am. Soc. Mass Spectrom. **9**, 1117 (1998).

[201] G. K. Schenter, S. M. Kathmann, and B. C. Garrett, J. Chem. Phys. **110**, 7951 (1999).

[202] Y. Ming, G. Lai, C. Tong, R. H. Wood, and D. J. Doren, J. Chem. Phys. **121**, 773 (2004).

[203] J. B. Watney, A. V. Soudackov, K. F. Wong, and S. Hammes-Schiffer, Chem. Phys. Lett. **418**, 268 (2006).

[204] C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models* (John Wiley & Sons. Ltd,).

[205] E. Wigner, J. Chem. Phys. **5**, 720 (1937).

[206] E. Wigner, Trans. Faraday Soc. **34**, 29 (1938).

[207] D. G. Truhlar and B. C. Garrett, Annu. Rev. Phys. Chem. **35**, 159 (1984).

[208] S. M. Kathmann, B. J. Palmer, G. K. Schenter, and B. C. Garrett, J. Chem. Phys. **128**, 064306 (2008).

[209] P. G. Bolhuis, C. Dellago, and D. Chandler, Proc. Natl. Acad. Sci. (U.S.A.) **97**, 5877 (2000).

[210] J. Gao, Acc. Chem. Res. **29**, 298 (1996).

[211] D. G. Truhlar et al., Acc. Chem. Res. **35**, 341 (2002).

[212] G. Cicotti, M. Gerrario, J. T. Hynes, and R. Kapral, J. Chem. Phys. **93**, 7137 (1990).

[213] A. Thomas, D. Jorand, C. Bret, P. Amara, and M. J. Field, J. Am. Chem. Soc. **121**, 9693 (1999).

[214] G. K. Schenter, B. C. Garrett, and D. G. Truhlar, J. Chem. Phys. **119**, 5828 (2003).

[215] J. Gao et al., Chem. Rev. **106**, 3188 (2006).

[216] T. C. Beutler, T.Bremi, R. R. Ernst, and W. F. van Gunsteren, J. Phys. Chem. **100**, 2637 (1996).

[217] J. Wang, Y. Deng, and B. Roux, Biophys. J. **91**, 2798 (2006).

[218] S. Boresch, F. Tettinger, M. Leitgeb, and M. Karplus, J. Phys. Chem. B **107**, 9535 (2003).

[219] B. Roux, M. Nina, R. Pomés, and J. C. Smith, Biophys. J. **71**, 670 (1996).

[220] J. Hermans and L. Wang, J. Am. Chem. Soc. **119**, 2707 (1997).

[221] D. A. Case et al., Amber 10, 2008.

[222] R. J. Loncharich, B. R. Brooks, and R. W. Pastor, Biopolymers **32**, 523 (1992).

[223] N. Castillo, P.-N. Roy, and T. L. Lowary, private communication.