Quantitative Studies in Effects of Additives on Protein Aggregation

By

Chetan Shinde

Submitted to the Department of Materials Science and Engineering in partial fulfillment of the requirements for the degree of Master of Science in Materials Science and Engineering at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2007

© Massachusetts Institute of Technology 2007. All rights reserved.

Author

Department of Materials Science and Engineering
May 25, 2007

Certified by

Bernhardt L. Trout
Associate Professor of Chemical Engineering
Thesis Supervisor

Certified by

Samuel M. Allen
Posco Professor of Physical Metallurgy
Thesis Reader

Accepted by

Samuel M. Allen
Posco Professor of Physical Metallurgy
Chair, Departmental Committee on Graduate Students
Quantitative Studies in Effects of Additives on Protein Stabilization

by

Chetan Shinde

Submitted to the Department of Materials Science and Engineering on May 25th, 2007 in partial fulfillment of the requirements for the degree of Master of Science in Materials Science and Engineering

Abstract

Rational design of protein additives has been limited by the understanding of mechanism of protein and additive interaction. In this work we have applied molecular dynamics with all atom potentials in order to study the thermodynamic effect of additives on proteins. The method is based on statistical mechanical model that characterizes the preferential binding of proteins to either water or additives. Extensive study was done on model systems comprising of additives urea, glycerol & arginine hydrochloride and proteins RNaseT1 and hen egg lysozyme. Trajectories in range 10-19 nanoseconds were analyzed in order to validate this method and compared with the experimental results. The method was found to agree with experimental results for the first 2 nanoseconds and the extended runs were studied further to narrow down the cause of deviations. Protein RNaseT1 was found to be very unstable and consequently showed very high deviations in preferential binding for longer runs. Constraining the protein using harmonic potential has resulted in better averages for RNase T1. Lysozyme has been found to be very stable and the calculations are in good agreement with experimental values. Local preferential binding calculations showed the importance of structure as well as sequence in prediction of preferential binding of protein.

Thesis Supervisor: Bernhardt Trout
Title: Associate Professor of Chemical Engineering
Acknowledgements

I would foremost like to thank my advisor, Prof. Bernhardt Trout, for giving me the opportunity to conduct research with him and provide me with valuable guidance. His sound advice, creative ideas, constant encouragement was an invaluable inspiration for my thesis.

A special thanks to Dr. Naresh for enormous help in my research. Discussions with him have always led to some new insights in the project, and made me a better researcher than I would have been had I not had the opportunity to work with him. I am also grateful to other members of the group: Bin Pan for providing help at crucial points when my codes crashed and inspiring with his ‘pony’ stories, Dr. Baynes my predecessor in the group working on protein stabilization, Gregg Beckham for his excellent maintenance of computers and all advice, Dr. Chong for helping me with CHARMM in initial stages. Outside of research I would like to thank Srikanth for interesting coffee-time chats, Jie for organizing lab outings and all the other people that I met during my short stay here, who have changed the way I look at things.

I am also indebted to Singapore MIT Alliance Flagship Research Program and Department of Materials Science and Engineering for financial support during my studies. Lastly, I acknowledge Prof. Samuel Allen for volunteering to be the thesis reader.
List of Figures

Figure 4.1 Snapshot of a protein in a mixed solvent box. The ribbon-like structure is the protein (RNase T1) and small dots in red color are water molecules and the additive (arginine) molecules are depicted in yellow. ........................................................................................................ 29

Figure 4.2 Plot showing the change in differential volume as a function of distance from protein surface for protein RNase T1 compared with the differential volumes for cylindrical and spherical shapes........................................................................................................ 31

Figure 5.1 Radial distribution functions for three systems averaged over first 2 ns. A) RNase T1 with water and urea B) RNaseT1 with water and glycerol C) RNaseT1 with water and arginine........................................................................................................ 38

Figure 5.2 $\tau_{23}$ plotted as a function of $r$, the distance from protein surface separating local from bulk domain. The graphs are average values obtained for first 2 ns for following systems A. RNase T1 in water and urea B. RNaseT1 in water and glycerol C. RNaseT1 in water and arginine........................................................................................................ 40

Figure 5.3 Time variation in the preferential binding coefficient with block averages over 200 nanoseconds for system RNase T1 with water and urea........................................ 43

Figure 5.4 Time variation in the preferential binding coefficient with block averages over 200 nanoseconds for system RNase T1 with water and glycerol ........................................ 43

Figure 5.5 Time variation in the preferential binding coefficient with block averages over 200 nanoseconds for systems RNase T1 with water and arginine ............................... 43

Figure 5.6 Time variation in the preferential binding coefficient with block averages over 200 nanoseconds for system lysozyme with water and urea. ................................. 44

Figure 5.7 Time variation in the preferential binding coefficient with block averages over 200 nanoseconds for system RNase T1 with water and urea........................................ 44

Figure 5.8 Time variation in the preferential binding coefficient with block averages over 200 nanoseconds for system with RNase T1 with water and urea, with the protein harmonically constrained................................................................. 45

Figure 5.9 Local preferential binding parameter for system: RNase T1 in urea and water; top figure represents averages for first 2ns and the bottom figure represents averages for entire run. The number of water molecules co-ordinated is plotted against the number of urea molecules co-ordinated for each contributing group. The dark line is the concentration line for the bulk. The location of a group above the concentration line.
shows a preferential binding to urea as compared to water and vice-a-versa. For the backbone showed by symbol B, the number of urea and water molecules were divided by ten to keep the axes in lower range so as to preserve details for amino acid groups.

Figure 5.10 Local preferential binding parameter of Serine residues for system: RNase T1 in urea and water. The top figure shows averages for first 2 ns and the bottom figure shows averages for entire run. The number of water molecules co-ordinated is plotted against the number of urea molecules co-ordinated for residue Serine, along with error bars. The dark line is the concentration line for the bulk. Serine residues have been labeled according to their number in protein sequence.

Figure 5.11 Local binding behavior of urea and water with the amino acid backbone and side chains in RNase T1: obtained from previous simulations (4). The labels are the one-letter code for each amino acid side chain, and "B" is the protein backbone. The line denotes the bulk urea concentration. In addition to the protein backbone and Ser, the hydrophobic amino acids Cys, Gly, Leu, Phe, Pro, Tyr, and Val all preferentially bind urea, while the hydrophilic Asp preferentially binds water.

Figure 5.12 Local preferential binding parameter for system: RNase T1 in glycerol and water; top figure represents averages for first 2 ns and the bottom figure represents averages for entire run. The number of water molecules co-ordinated is plotted against the number of glycerol molecules co-ordinated for each contributing group. The dark line is the concentration line for the bulk. The location of a group above the concentration line shows a preferential binding to glycerol as compared to water and vice-a-versa. For the backbone showed by symbol B, the number of glycerol and water molecules were divided by ten to keep the axes in lower range so as to preserve details for amino acid groups.

Figure 5.13 Plot of ratio \( p(\tau_b) = \frac{\tau_b\sigma^2(<A_1>)}{\sigma^2(A)} \) against \( \tau_b^{1/2} \) for MD simulation run of RNase T1 in urea. In this plot the quantity \( A \) is preferential binding parameter. The statistical efficiency \( s \) is the value at which the plot goes to a constant plateau which is shown by a dotted line. The dotted line shows the value where a plateau is approached.

Figure 5.14 Plot of ratio \( p(\tau_b) = \frac{\tau_b\sigma^2(<A_1>)}{\sigma^2(A)} \) against \( \tau_b^{1/2} \) for MD simulation run of RNase T1 in urea. In this plot the quantity \( A \) is local preferential binding parameter for backbone. The dotted line shows the value where a plateau is approached.

Figure 5.15 Plot of ratio \( p(\tau_b) = \frac{\tau_b\sigma^2(<A_1>)}{\sigma^2(A)} \) against \( \tau_b^{1/2} \) for MD simulation run of lysozyme in urea. In this plot the quantity \( A \) is local preferential binding parameter for protein. The dotted line shows the value where a plateau is approached.
Figure 5.16 The RMSD (in units of Å) results for RNaseT1 in water and solvents: urea, glycerol and arginine. The fourth system is protein in water without cosolvent. The protein was recentered and reoriented for these calculations.............................................. 57

Figure 5.17 The RMSD (in units of Å) results for Lysozyme in water and solvents: urea, glycerol and arginine. The protein was recentered and reoriented for these calculations. ........................................................................................................................................... 57

List of Tables

Table 4.1 List of the systems under study; n1 stands for number of water molecules, n3 stands for number of cosolvent molecules. ........................................................................................................................................... 28

Table 5.1 Cosolvent Properties................................................................................................................. 37

Table 5.2 Table showing Preferential Binding Parameter obtained from 1. MD simulations average for first 2 ns, 2. MD simulations average for entire run, 3. Experimental values extrapolated to the concentration of interest (34) (35) (36) (37). (Values marked * were obtained from experiments and analysis performed by Curt Schneider in Trout Lab using Vapor Pressure Osmometry) ............................................................................................................. 41

Table 5.3 Number of water and urea molecules co-ordinated for each serine residue identified by its number in sequence. The numbers represent averaging over the entire run ........................................................................................................................................... 50
1 Introduction

1.1 Background

A recent survey listed 324 biotechnology medicines, either in human clinical trials or under review by regulatory agencies. These biotechnology medicines cover nearly 150 diseases, including cancer, infectious diseases, autoimmune diseases and AIDS/HIV. The increased use of recombinant DNA technique for therapeutic proteins in the pharmaceutical industry has highlighted issues like protein stability and efficacious delivery without side effects (1). Proteins are useful in therapeutics because they have a wide range of physiological functions and are extremely potent. Their therapeutic activity is highly dependent on their conformations. However, the protein structure is very flexible and sensitive to external conditions, which means that production, formulation and handling of proteins needs to be done very carefully. Otherwise, it can lead to various physical and chemical pathways of deterioration which include aggregation, deamidation, oxidation, and hydrolysis. Of these pathways protein aggregation is arguably the most common and troubling manifestation of protein instability, almost in all phases of protein drug development.

Protein aggregation also plays a significant role in human etiology and has been attributed to at least 20 different diseases (2). Among these are Alzheimer’s disease, Parkinson disease, prion diseases (bovine spongiform encephalopathy and Creutzfeldt-Jacob diseases), Huntington’s disease, Down’s syndrome, cataract, and sickle cell disease (3). This has underscored the need to develop an understanding of the entire aggregation process. Therefore, achieving a better understanding of protein aggregation is critical not only in various biopharmaceutical processes but also in finding a solution to those devastating diseases. There has been a lot of effort to study this problem of aggregation, especially in terms of protein interaction with mixed solvent.
Proteins are seldom solvated in pure water. Other solvent components such as buffer salts and stabilizers are always present in the laboratory or in formulations of therapeutic proteins. Even in intracellular solutions there are numerous other chemical species like metabolites, nucleic acids, osmolytes and other molecules. These other components that are called “cosolvents” affect the protein chemical potential and consequently its tendency to form aggregates or to fold in the native state (4). Experimental results have confirmed that adding low molecular weight components, such as salts, sugars, or polyols to protein solution can effectively prevent aggregation.

However, the phenomenon of aggregation and the way additives affect proteins is still an unsolved problem. Every protein has a different sequence and structure, necessitating the use of a different additive. There is often no theoretical guidance to aid in selection of optimal additives and stabilization is mostly restricted to heuristic experimental screens.

1.2 Objectives and Overview

With this background, the aim of this thesis work is to give a detailed description of protein and cosolvent interaction in order to understand the mechanism behind the effects of additives on protein aggregation. The focus will be on using a preferential binding parameter, the calculation of which is based on statistical mechanical techniques without any use of adjustable parameters to characterize binding of additives. The model systems chosen to study these effects are proteins - RNase T1 & hen egg Lysozyme and additives - urea, glycerol & arginine. This understanding will help us in setting up an algorithm for rational design of protein additives.

The thesis is organized in five chapters. Chapter 2, introduces to the readers the problem of protein aggregation. This introduction deals with basics of protein structure, the relation between folding and aggregation and the mechanisms for protein aggregation.
Additives play an important part as denaturants or stabilizing agents for proteins. Chapter 3 deals with more formal thermodynamical treatment of the effect of additives on protein chemical potential. This chapter defines terms like transfer free energy, preferential interaction parameter and preferential binding parameter. Since, the preferential binding parameter calculation is the focus of this thesis; this chapter also gives the physical interpretation of this parameter.

Chapter 4 contains the details of molecular dynamics simulations performed to study the systems of interest. It also outlines the basic algorithm used to calculate the preferential binding parameter for a protein as well as its constituent groups. The error analysis part covers the analysis of variances in equilibrium properties taking into consideration the correlated steps.

The results and discussions form the major part of Chapter 5. A detailed description of results obtained for the preferential binding parameter is given by comparing with experimental values and previous simulations. Chapter 6 summarizes the findings with major conclusions and suggestions for future work.
2 Protein Aggregation

2.1 Structure of Proteins

The word *protein* comes from the Greek word "proteis", meaning "of primary importance" and were first described and named by Berzelius in 1838. However, their central role in living organisms was not fully appreciated until 1926, when James B. Sumner showed that the enzyme urease was a protein (5). The first protein to be sequenced was insulin. True to the Greek meaning of protein, they play crucial life-sustaining biological roles, both as constituent molecules and as triggers of physiological processes for all living organisms.

Protein molecules come in a wide range of sizes and functionalities. Proteins are linear polymers built from 20 different L-α-amino acids. All amino acids share common structural features including an α carbon to which an amino group, a carboxyl group, and a variable side chain are bonded. The side chains of the standard amino acids, have different chemical properties that produce proteins' three-dimensional structure and are therefore critical to protein function. The amino acids in a polypeptide chain are linked by peptide bonds formed in a dehydration reaction. Once linked in the protein chain, an individual amino acid is called a residue and the linked series of carbon, nitrogen, and oxygen atoms are known as the main chain or protein backbone. The peptide bond has two resonance forms that contribute some double bond character and inhibit rotation around its axis, so that the alpha carbons are roughly coplanar. The other two dihedral angles in the peptide bond determine the local shape assumed by the protein backbone.

What differentiates proteins other than number of amino acids is the sequence of amino acids. Depending on their sequence proteins fold into a 3-dimensional structure known as
its native structure. Four distinct aspects of protein's structure are: Primary, Secondary, Tertiary and Quaternary (6).

The ability of the reduced and unfolded protein to spontaneously fold into its native state established that the primary amino acid sequence of a protein contains all of the information necessary for proper folding into native form, a fundamental principle for which Anfinsen received the Nobel Prize in Chemistry in 1972. Despite the considerable effort, to gain understanding of the fundamentals of folding, we are still not able to give a detailed description of the mechanism by which any protein folds.

2.2 Protein Aggregation

Aggregation is a ubiquitous protein stabilization problem because aggregation is related to the natural process of protein folding. The driving force for protein folding is free energy minimization and the process of folding can be described thermodynamically as the finding of minima on an energy landscape that is of the shape of a funnel. The same free energy minimization force drives aggregation. Protein folding is possible because all the bonds in a polypeptide chain, with the exception of the peptide bonds and the bonds of aromatic rings can rotate freely. The conformation of a given folded chain is highly specific to endow the protein with its biological functions, but it is only marginally stable at room temperature. Both these properties result from the fact that the conformation is stabilized by many weak, non-covalent interactions involving both main-chain and side-chain atoms. For these reasons, protein folding is subject to errors, described by the terms misfolding and aggregation.

Nonnative protein aggregation describes the assembly from initially native, folded proteins of aggregates containing nonnative protein structures. Aggregation is often irreversible, and aggregates often contain high levels of nonnative, intermolecular β-sheet structures. Protein molecules may aggregate simply by physical association with one another without any changes in primary structure (physical aggregation) or by formation
of a new covalent bond(s) (chemical aggregation). Formation of such a bond(s) can either directly crosslink proteins (aggregation), or indirectly alter the aggregation tendency of the original protein (7). Changing protein’s environmental conditions often lead to changes in protein aggregation behavior and therefore it is important to study how different solute conditions affect protein stability. Protein aggregation behaviors, such as onset, aggregation rate, and the final morphology of the aggregated state (i.e., amorphous precipitates or fibrils) have been found to depend strongly on the properties of a protein’s solution environment, such as temperature, pH, salt type, salt concentration, cosolvents, preservatives, and surfactants, as well as the relative intrinsic thermodynamic stability of the native state (8).

A major driving force for both aggregation and folding is the reduction of exposure area of hydrophobic side chains. This internalization of hydrophobic chains and exposure of hydrophilic chains lowers the free energy of the protein. Synthesis conditions or some of the above environmental factors may influence this stable conformation of protein and cause exposure of hydrophobic groups. The reduction of hydrophobic exposure may come in the form of intermolecular association to form non-functional aggregates.

### 2.3 Aggregation Mechanisms

#### 2.3.1 Folding/Unfolding Intermediates

There is overwhelming evidence of the presence of an intermediate state between unfolded state and aggregates called the intermediate state. These intermediate states are very unstable as opposed to stable native state or even unfolded state. A contiguous hydrophobic patch is necessary to initiate aggregation, which is why folded proteins (with buried hydrophobic groups) and unfolded ones with random hydrophobic groups are less prone to aggregation. It has also been proposed that higher folding barriers help prevent aggregation (9) (10). The aggregation process can be described by equation 2.1
where proteins form reversible unfolding intermediates (I) from native state (N), which then form reversible unfolded proteins (U) or irreversible/reversible aggregates (A).

\[
\text{U} \leftrightarrow \text{I} \leftrightarrow \text{N} \leftrightarrow \text{A}
\]  

(2.1)

Further growth of protein aggregates can take place by monomer-cluster growth (monomer adds to a growing multimer) and cluster-cluster growth (a multimer adds to another multimer) (11) (12).

### 2.3.2 Denatured-State Aggregation

This model of aggregation is based on reaction 2.2, where denatured proteins aggregate directly and not through any intermediate state. This model is supported by experimental studies on apomyoglobin by De Young et al. and protein solubility models studied by Arakawa and Timasheff (13) (14).

\[
\text{N} \leftrightarrow \text{D} \leftrightarrow \text{A}
\]  

(2.2)

The first equilibrium between denatured state (D) and native state (N) is a balance between conformational entropies that forces the chains to open and the hydrophobic force that forces the chain to fold. The hydrophobic force also drives aggregation in
second equilibrium, where it is more favorable to have inter-molecular hydrophobic contacts as opposed to intra-molecular hydrophobic contacts due to more conformational freedom of chains in aggregated states (11) (15).
3 Effect of Additives on Proteins

The tendency of proteins to aggregate causes grave problems in biotechnological and the pharmaceutical industry where they are synthesized, processed and stored at very high concentrations. These aggregated proteins may lose their biological activity, can often be immunogenic, and can also have acute toxic effects in vivo. In order to counter aggregation there have been efforts on various levels, which include substitution and chemical modification of protein, or the controlling protein environment by additives. Empirically it has been observed that addition of low molecular weight components like salts, sugar or polyols to protein solution results in change in aggregation equilibrium (3). The effects of the presence of these components, called ‘cosolvents’, will be discussed in this chapter.

3.1 Thermodynamics of Cosolvents

In presence of cosolvents the chemical equilibrium for the aggregation reactions is altered because cosolvents have a different stabilizing effect on each of initial, intermediate and final states. Cosolvents give rise to changes in experimentally observable quantities such as equilibrium constants and reaction rates. Most importantly, the aggregation rate can change. To understand how additives affect aggregation, we must understand how they affect the free energy barrier of the rate-limiting step in the aggregation process. The effects can be described by three related thermodynamic parameters and their changes during the course of reaction, namely the transfer free energy, the preferential interaction parameter and the preferential binding parameter.
3.1.1 The Transfer Free Energy

This term captures the change in interaction energy of protein with solvent when it is transferred from pure water to a cosolvent system. In Scatchard notation (16), where water, protein and cosolvent are designated as component 1, 2 and 3 respectively the transfer energy \( \Delta \mu^r_2 \) is given by:

\[
\Delta \mu^r_2 = \mu_2^{(\text{cosolvent})} - \mu_2^{(\text{water})}
\]  

(3.1)

Where \( \mu_i \) is the chemical potential of component i. This equation is applicable to any state of protein along the reaction co-ordinate and the knowledge of \( \Delta \mu^r_2 \) will give us the value of \( \mu_2 \).

3.1.2 Preferential Interaction Parameter

The preferential interaction parameter captures the mutual perturbations of the chemical potentials of the protein and cosolvent. If we examine the situation in a reciprocal manner; when a protein is added to a water-cosolvent mixture, it interacts with the potential of the cosolvent i.e. the chemical potential of the cosolvent is disturbed by the protein, \( \left( \frac{\partial \mu_2}{\partial m_3} \right)_{T,P,m_0} \) and that of the protein by cosolvent \( \left( \frac{\partial \mu_3}{\partial m_2} \right)_{T,P,m_0} \) (17).

Here \( m \) stands for concentration in terms of molality, thus the first term gives the gradient in the transfer free energy with respect to concentration of cosolvent. The total change in chemical potential of protein due to transfer to a cosolvent solution is given by the following integral:

\[
\Delta \mu^r_2 = m \left( \frac{\partial \mu_2}{\partial m_3} \right)_{T,P,m_0} dm_3
\]  

(3.2)

Where T and P have their usual meaning of temperature and pressure.
3.1.3 Preferential Binding Parameter

This is the manifestation of the perturbation of chemical potential of the protein and it can be experimentally measured. The abovementioned mutual perturbations lead to redistribution of solvent components in the vicinity of proteins. The expression is as follows:

\[
\left( \frac{\partial m_3}{\partial m_2} \right)_{T,P,\mu_3} = -\frac{(\partial \mu_2/\partial m_3)_{T,P,m_3}}{(\partial \mu_3/\partial m_3)_{T,P,m_2}}
\]  

(3.3)

The denominator denotes the cosolvent non-ideality. With a small approximation that, 
\[(\partial m_3/\partial m_2)_{T,P,\mu_5} = (\partial m_3/\partial m_2)_{T,P,\mu_5},\]
we can measure this quantity experimentally at dialysis equilibrium. This preferential binding parameter is denoted by \(\bar{V}\) in Scatchard notation (16) and by \(\Gamma_{23}\) in Cassassa and Eisenberg (18) and Schellman (19) notation. Henceforth, \(\Gamma_{23}\) will be used to denote preferential binding parameter.

The thermodynamical relation between preferential binding parameter \((\Gamma_{23})\) and change in chemical potential of component 3 is (20):

\[
\left( \frac{\partial m_3}{\partial m_2} \right)_{T,P,\mu_5} = -\frac{(\partial \mu_2/\partial m_3)_{T,P,m_3}}{(\partial \mu_3/\partial m_3)_{T,P,m_2}}
\]  

(3.4)

Thus the left hand side of equation (3.2) is modified as the integral of two terms given by:

\[
\Delta \mu_2^{ir} = -m_3 \int_0^{m_3} \left( \frac{\partial \mu_2}{\partial m_2} \right)_{T,P,m_3} \left( \frac{\partial m_2}{\partial m_3} \right)_{T,P,\mu_3} \partial m_3
\]  

(3.5)

\[
= -m_3 \int_0^{m_3} \left( \frac{\partial \mu_2}{\partial m_2} \right)_{T,P,m_3} \Gamma_{23} \partial m_3
\]  

(3.5)
The first term inside the integral captures the gradient of chemical potential of the protein with change in concentration of cosolvent and can be experimentally evaluated on an additive mixture with water at very dilute concentrations \((m_0 \to 0)\).

### 3.2 Preferential Binding Parameter: Physical Interpretation

Preferential binding parameter measures the excess number of cosolvent molecules in the vicinity of protein molecule (component 2) as compared to the bulk solvent. When \(\left(\frac{\partial \mu_j}{\partial m_2}\right)_{T,p,m_3}\) is negative, i.e., the interaction is favorable, from equation 3.5 we have \(\Gamma_{23}\) as positive, and vice-a-versa. Schellman (19) and Kirkwood and Goldberg (21) have shown the connection between the thermodynamic definition and the intuitive notion of binding, based on statistical mechanics:

\[
\Gamma_{23} = \left\langle n_3^\mu - n_2^\mu \left(\frac{n_3^j}{n_1^j}\right)\right\rangle 
\]

(3.6)

Where \(n_j^i\) denotes the number of molecules of species \(j\) in domain \(i\) and the angled brackets \(< >\) stand for ensemble average. Subscripts 1, 2 and 3 stand for water, protein and cosolvent respectively. Superscripts I and II stand for bulk and local domain respectively. When the cosolvent concentration is higher in the local domain of the protein as compared to the bulk domain, \(\Gamma_{23}\) is positive and \(\Delta \mu_j^\sigma\) is negative indicating a favorable interaction. On the other hand, a lower cosolvent concentration in the vicinity of protein leads to negative \(\Gamma_{23}\) and positive \(\Delta \mu_j^\sigma\), indicating an unfavorable interaction.

Thus, the modifier "preferential" essentially indicates that the protein has higher affinity ("preference") for one solvent over other. The way in which cosolvents affect any of the aggregation mechanisms mentioned earlier, depends on the balance between the transfer free energies of the protein from water to the cosolvent system in the two end and
intermediate states of the reaction. The value of $\Delta \mu^\nu_2$ in either of the states may be positive or negative, depending on whether $\Gamma_{23}$ is negative or positive respectively.
4 Methodology

The calculation of preferential binding using the ensemble average equation explained in the previous chapter requires the measurement of the number of cosolvents and water molecules in local and bulk domain. Molecular Dynamics was used in this work to create an equilibrium ensemble of protein in a mixed solvent system; and ensemble averages were calculated from this system.

4.1 Molecular Dynamics

Molecular Dynamics (MD) is an atomistic simulation method characterized by treatment of every atom by a point mass and integration of Newton equations to advance the atomic positions and velocities. The classical equations of motion for solute and solvent atoms are treated explicitly and integrated numerically. The initial positions of atoms are usually determined from X-ray or NMR structures and initial velocities are assigned using Maxwell distribution at some temperature near zero. The temperature is then increased to the desired temperature by scaling the velocities of all atoms. The system is then equilibrated to prevent localized increase in energy persisting throughout the simulations. This equilibrated trajectory is then run for extended times to analyze equilibrium properties.

The molecular dynamics package used for this work was version 31 of CHARMM (22). We utilize explicit atomic interaction potentials (force fields), such as Lennard-Jones, Coulombic, spring, and torsion interactions, with pre-fit coefficients. CHARMM force field was used to compute forces and an explicit solvent model was used with the TIP3P model for water (23). The potentials used for urea and glycerol were obtained using standard CHARMM geometries, partial charges and parameters used in previous works (4) (22) (24) (25). Potentials and partial charges for L-α-arginine and guanidinium ions were obtained from the standard CHARMM potential for arginine and guanidinium respectively (22). Counter-ions of chloride were used to balance charges for both
additives. Proteins used in this study, RNase T1 (PDB: 1ygw) and Hen Lysozyme (PDB: 1e8l) were obtained from the Protein Data Bank (26). The pH was fixed at 7 for all simulations, which was achieved by setting the protonation states of amino acid side chains in the appropriate form. Arginine, cysteine, lysine, and tyrosine were protonated while Aspartate, glutamate, and histidine with pKa values of 3.4, 4.1, and 6.6 (27) respectively, were deprotonated at pH 7.

Periodic boundary condition was used with a truncated octahedron box for simulation of the protein in a mixed solvent system. The specific shape of the truncated octahedron was selected because of the spherical nature of proteins under study. The electrostatic interactions were calculated using the particle mesh Ewald summation. The size of the box was selected so as to have a shell of solvent of at least 10Å from the surface of protein. The numbers of water and cosolvent molecules in the box were chosen to keep the molal concentration at 1m (18 molecules of cosolvent for each 1000 molecules of water); they are listed in Table 4.1. The required concentration of water and cosolvent and the corresponding counter-ions were randomly placed inside the box, after which it was minimized at 0K and heated to 298K followed by equilibration for 100 picoseconds. Protein was then introduced at the center of the box and overlapping molecules were deleted. The counter-ions for proteins were placed using SOLVATE 1.0. This system was minimized at 0K, after which it was heated to 298K, followed by equilibration at same temperature and pressure of 1 atmosphere. The equilibrated system was then used to get trajectories at constant pressure and temperatures of varying lengths with time steps of 2 femto-second. The co-ordinates from this trajectory are saved at every 0.1 picoseconds and this data is analyzed for calculating thermodynamical properties.

Table 4.1 List of the systems under study; n1 stands for number of water molecules, n3 stands for number of cosolvent molecules.

<table>
<thead>
<tr>
<th>System No.</th>
<th>Protein</th>
<th>Additive</th>
<th>n1</th>
<th>n2</th>
<th>molality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RNaseT1</td>
<td>Urea</td>
<td>4544</td>
<td>90</td>
<td>1.1m</td>
</tr>
<tr>
<td>2</td>
<td>RNaseT1</td>
<td>Glycerol</td>
<td>4596</td>
<td>87</td>
<td>1.1m</td>
</tr>
<tr>
<td>3</td>
<td>RNaseT1</td>
<td>Arg⁺</td>
<td>4110</td>
<td>90</td>
<td>1.2m</td>
</tr>
<tr>
<td>4</td>
<td>Lysozyme</td>
<td>Urea</td>
<td>8353</td>
<td>157</td>
<td>0.9m</td>
</tr>
<tr>
<td>5</td>
<td>Lysozyme</td>
<td>Glycerol</td>
<td>7538</td>
<td>157</td>
<td>1.0m</td>
</tr>
<tr>
<td>6</td>
<td>Lysozyme</td>
<td>Arg⁺</td>
<td>7990</td>
<td>154</td>
<td>0.9m</td>
</tr>
</tbody>
</table>
Figure 4.1 Snapshot of a protein in a mixed solvent box. The ribbon-like structure is the protein (RNase T1) and small dots in red color are water molecules and the additive (arginine) molecules are depicted in yellow.

4.2 Calculation of Preferential Binding Parameter

The method of calculating the preferential binding parameter, based on a statistical mechanical method applied to all-atom model with no adjustable parameter, was developed by Baynes and Trout (4). The approach is used to calculate number of ‘bound’ molecules to the protein without \textit{a priori} information of any ‘binding sites’ on the protein. Since the method is based on molecular level approach, we get a detailed description of interactions between protein and cosolvents. The variation of concentration as a function of distance from protein surface can be computed. Similarly, we can calculate $\Gamma_{23}$ as a function of distance from protein and identify the distance where it approaches a constant
value. Moreover, it is possible to find out $\Gamma_{23}$ and $\Delta\mu_2''$ of systems where no experimental data is available or possible, such as transition state configurations or unstable states of proteins. Another advantage of this method is that the calculation of $\Gamma_{23}$ or $\Delta\mu_2''$ requires just one trajectory as opposed to 15-20 trajectories required for thermodynamic integration (28) (29). This method of calculation will be validated using abovementioned systems. The comparison with experimental values and previous simulations will help us fine tune the method.

The MD run obtained from CHARMM, is saved at periodic time intervals and these saved frames are used to find several properties one of which is $\Gamma_{23}$. The following points elucidate the algorithm used in calculation of the preferential binding parameter:

1. Every molecule (water and cosolvent atoms) is treated as a point at its center of mass.
2. Its distance from the surface of all protein atoms is calculated. Here, surface of atom is defined as the sphere with Van der Waal's radius.
3. The minimum of all such distances is identified and is put in bins of size $0.1\,\text{Å}$. It has been found that accuracy of $0.1\,\text{Å}$ is required to capture details in the variation of $\Gamma_{23}$ with distance.
4. Thus each molecule is associated with some distance and a number density function is obtained as a function or distance $r$ from protein's surface: $\rho_1(r)$ for water and $\rho_3(r)$ for cosolvent. For a hypothetical case of spherical protein the dividing factor in this case would be proportional to $r^2$, and $r$ for cylindrical protein. However, protein in this case is much simpler and the dividing factor in this case is a complicated function of $r$ and is shown in Figure 4.2.
5. The region in which the number density function goes to a constant value is identified as bulk and the bulk number density $\rho(\infty)$ is used to find the radial distribution function using the formula:

$$g(r) = \frac{\rho(r)}{\rho(\infty)} \quad (4.1)$$
where $r$ is the time period of entire run and $\Gamma_{23}(t_i)$ stands for the value of preferential binding coefficient at time $t_i$. This value can be calculated by another method using radial distribution functions for water and cosolvent defined in equation 4.1,

$$\Gamma_{23}(t) = n^H_3(t) - n^H_2(t) \left( \frac{n^L_3(t)}{n^L_1(t)} \right)$$

$$= \rho_3(\infty) \int g_3(r) dV - \left( \frac{\rho_3(\infty)}{\rho_1(\infty)} \right) \rho_1(\infty) \int g_1(r) dV$$

where the integral extends from $r = 0$ to $\infty$; it should be noted that the expression inside the integral is equal to zero inside the bulk domain. Let $r^*$ be the distance from the surface of protein at which bulk domain begins. It should be noted that $r^*$ should be sufficiently away from protein for $g_3(r)$ and $g_1(r)$ to be equal to 1. The value of $g_3(r)$ and $g_1(r)$ should be used to define the value of $r^*$. Since the box size is limited for MD simulations, this integral is evaluated from protein surface to the box boundary ($\sim 10\AA$). It will be shown in results section that the integral attains a constant value within the range of 6-8Å.

**4.3 Calculation of local preferential binding parameter**

The protein surface is heterogeneous and made up of various constituents with varying preference for cosolvent and water. Thus the local concentration of cosolvent and water may be different and may depend on the nature of the group. The preferential binding parameter for the protein gives the total preferential binding for the entire protein and may not be same for each constituent groups. The total interaction of mixed solvent with protein must be conceived as a large number of small interactions involving every group that makes direct contact with the solvent (19).
The distance $r^*$ is defined as distance at which there is no significant difference between $\rho(r)$ and $\rho(\infty)$. The molecules with centre of mass inside $r^*$ are said to be belong to the local domain (II) while those outside are said to belong to the bulk domain.

![Figure 4.2 Plot showing the change in differential volume as a function of distance from protein surface for protein RNase T1 compared with the differential volumes for cylindrical and spherical shapes.](image)

6. Before we come to the step of calculating the preferential binding coefficient, we need to review the expression obtained from statistical analysis. From equation 3.6 we define instantaneous $\Gamma_{23}(t)$ as:

$$
\Gamma_{23}(t) = n_2^H(t) - n_2^H(t) \left( \frac{n_3'(t)}{n_1'(t)} \right)
$$

(4.2)

For each time instance in trajectory and the preferential binding for the entire trajectory is defined as the time average of all these instantaneous values:

$$
\Gamma_{23} = \frac{\sum_{i=0}^{\tau} \Gamma_{23}(t_i)}{\tau}
$$

(4.3)
As proposed by Tanford (30) a protein is considered as a set of non-overlapping constituent groups. The transfer free energy for the protein (defined by equation 3.1) is then a summation of contributions of various groups such as amino acid side chains and the protein backbone as shown in following expression (31):

$$\Delta \mu_{2}^{tr} = \sum_{i} \alpha_{i} \Delta g_{i}^{tr}$$ \hspace{1cm} (4.7)

where $\Delta g_{i}^{tr}$ is the transfer free energy for constituent model group and $\alpha_{i}$ is the solvent accessible area of the constituent in the protein, normalized to the solvent accessible area of the model group by itself. Similarly, it is possible to extend this group contribution theory for the preferential binding parameter. If we define $\Gamma_{23,j}^{r}$ as the contribution by a constituent group then the preferential binding parameter for the entire protein is given by (4):

$$\Gamma_{23} = \sum \Gamma_{23,j}^{r}$$ \hspace{1cm} (4.8)

Thus the overall preferential binding coefficient can be predicted if the sequence and structure of protein is known. The calculation of preferential binding parameter for constituent groups is similar to that of protein and is based on equation 3.6.

$$\Gamma_{23,i,d} = \left(n_2^{\alpha} - n_2^{\alpha} \left( \frac{n_3^{\alpha}}{n_3^{\alpha}} \right) \right)$$ \hspace{1cm} (4.9)

where the subscript $i$ denotes a particular contributing group so that $n_3^{\alpha}$ and $n_1^{\alpha}$ denote the number of cosolvent and water molecules in the local domain that are nearest to group $i$. The following steps elucidate the algorithm used in calculation group preferential binding coefficients for constituent groups:
1. The protein is divided into 21 constituent groups: the 20 amino acids and backbone. Here, the protein backbone is defined as the –NH–CH–CO– as well as the extra proton at the N-terminus and the extra OH at the C terminus.

2. As in the case of the preferential binding parameter for protein, we treat every molecule as a point at its center of mass. The nearest group, in terms of distance from the Van der Waal’s surface, to every molecule is found and the molecule is assigned to that particular group.

3. We find $n_{3d}^n$ and $n_{l}^n$ associated with every molecule and $\Gamma_{23d}$ can be computed.

4.4 Error Analysis

Computer simulations are subject to both systematic and statistical errors. If we can perform simulations ad infinitum then the data generated will on averaging give exact numbers to satisfy our model. However, simulation averages are taken over a run of finite duration and this leads to statistical imprecision in the mean values obtained.

According to Central Limit Theorem, as $N \to \infty$, the limiting distribution for a sum of random variables is the normal distribution. Thus a simulation run average can be thought of as sampling from some limiting Gaussian distribution about a true mean, since it is averaging over many steps. For a Gaussian distribution, all moments are determined by the first two, the mean and the variance. Therefore, in order to characterize a distribution for a quantity of interest such as $A$, it is sufficient to find $<A>$ and $<\delta A^2>$. For our long (but finite) simulations we need to compute averages and variances assuming that they obey Gaussian statistics approximately.

According to Law of Large Numbers, the average of $N$ sampled random variables converges (in probability) to its expected value. For simulation data that contains a total of $\tau_{run}$ time steps, the run average of some quantity $A$ is defined as:
\[ < A >_{\text{run}} = \frac{1}{\tau_{\text{run}}} \sum_{\tau=1}^{\tau_{\text{run}}} A(\tau) \]  

For statistically independent observations of \( A(\tau) \), according to analysis by Jacucci and Rahman (32), the variance in the mean is given by:

\[ \sigma^2(< A >_{\text{run}}) = \sigma^2(A) / \tau_{\text{run}} \]  

\[ \sigma^2(A) = \delta A^2 = \frac{1}{\tau_{\text{run}}} \sum_{\tau=1}^{\tau_{\text{run}}} (A(\tau) - < A >_{\text{run}})^2 \]  

However, the data points in our simulation are not independent. Therefore, the entire run is broken down into blocks of length \( \tau_b \), such that there are \( n_b \) such intervals and \( n_b \tau_b = \tau_{\text{run}} \). The mean for such blocks \( < A >_b \) can be calculated as:

\[ < A >_b = \frac{1}{\tau_b} \sum_{\tau=1}^{\tau_b} A(\tau) \]  

All such means for \( n_b \) blocks are used to calculate variance in means:

\[ \sigma^2(< A >_b) = \frac{1}{n_b} \sum_{b=1}^{n_b} (< A >_b - < A >_{\text{run}})^2 \]  

As the block length becomes large enough to be statistically uncorrelated, the variance in means of block averages is inversely proportional to block length. We need to find this constant of proportionality in order to evaluate the statistical error in our run. This constant is defined as (33):

\[ s = \lim_{\tau_b \to \infty} p(\tau_b) = \lim_{\tau_b \to \infty} \frac{\tau_b \sigma^2(< A >_b)}{\sigma^2(A)} \]  

\[ s = \lim_{\tau_b \to \infty} p(\tau_b) = \lim_{\tau_b \to \infty} \frac{\tau_b \sigma^2(< A >_b)}{\sigma^2(A)} \]
This method of block average is a powerful method to determine whether the simulation is long enough to yield a reliable estimate of a particular quantity. The quantity $s$ is called statistical inefficiency and any technique that reduces $s$ will help us to calculate more accurate simulation averages.
5 Results and Discussion

5.1 Radial Distribution Function

As outlined in equation 4.6, the radial distribution function is the intermediate step in calculation of the preferential binding parameter and it gives important clues about the protein environment in terms of relative concentrations of water and cosolvent molecules. The three cosolvent molecules used in separate systems: urea, glycerol and arginine have different sizes which have been characterized in Table 5.1. The radius of gyration was calculated using CHARMM and average radius was calculated from three principle diameters of cosolvent molecules, by averaging and dividing by two.

<table>
<thead>
<tr>
<th>Cosolvent</th>
<th>Chemical formula</th>
<th>Molecular Mass (g/mol)</th>
<th>Radius of gyration (Å)</th>
<th>Average radius (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urea</td>
<td>C$_3$H$_5$(OH)$_3$</td>
<td>60.1</td>
<td>1.4</td>
<td>1.66</td>
</tr>
<tr>
<td>Glycerol</td>
<td>(NH$_2$)$_2$CO</td>
<td>92.1</td>
<td>2.0</td>
<td>2.19</td>
</tr>
<tr>
<td>Arginine hydrochloride</td>
<td>C$<em>6$H$</em>{14}$N$_4$O$_2$·HCl</td>
<td>210.6</td>
<td>3.2</td>
<td>2.84</td>
</tr>
</tbody>
</table>

The radial distributions for each of these cosolvents and water with protein RNase T1 is plotted in Figure 5.1. As outlined in the method of calculation these were calculated from the protein’s surface and therefore the Van der Waals repulsion is not as high as would be expected if the function was plotted as a distance from atom centers. The molecules were treated as points at their center of mass, which results in these centers being closer to the surface. Moreover, the more dominant electrostatic forces play an important role in binding and hence pull the cosolvent molecules closer. As the size of molecule increases the repulsion also increases. A prominent first co-ordination shell is displayed at a distance which is closely related to the average radius of the molecule. Urea being the smallest of additives has its first co-ordination shell at 1.7Å followed by glycerol which
is at 2.3 Å. These molecules also show a less prominent second co-ordination shell. Arg$, which is a bigger molecule as compared to the other two, does not show a prominent peak, rather a plateau like region is displayed which extends from 2-3 Å. The water molecules corresponding to the system having arginine hydrochloride as an additive also show a slightly higher peak than that of water with urea and water with glycerol. The plot for Arg$ also appears to be highly undulated as compared to others. Being a larger molecule arginine has a lower diffusivity and does not translate through the box as fast as a small molecule and so in order to get a smoother curve it is necessary to take time average over a longer run. A closer study of this plot at higher resolution reveals the second co-ordination shell for arginine around 6 Å.

![Figure 5.1 Radial distribution functions for three systems averaged over first 2 ns. A) RNase T1 with water and urea B) RNase T1 with water and glycerol C) RNase T1 with water and arginine](image)

All plots go to an almost constant value after 6-7 Å, indicating the presence of the bulk domain after this distance. A notable feature in radial distributions is the magnitude of
co-ordination shell peaks. A higher peak does not necessarily imply more cosolvent in the vicinity of protein. Any effort to characterize preferential binding should be based on equation 4.6 that takes difference of radial distributions and integrates them over the distance.

5.2 Preferential Binding Parameter

The key to calculating the preferential binding parameter is estimating the distance \( r^* \), which separates the local and the bulk domain. This distance should not be so close to the protein surface that it lies in the local region, at the same time we cannot go too far as we are limited by the size of the box. In order to decide the optimum distance \( r^* \), we plot \( \Gamma_{23} \) as a function of distance \( r \) from protein surface. This quantity \( \Gamma_{23}(r) \) is called the apparent preferential binding coefficient and has been plotted in Figure 5.2 for system of RNase T1 and with three different additives. This plot helps us define \( r^* \) so that the error in the value of the preferential binding parameter is minimized.

The apparent preferential binding parameter at any given distance \( r \) from the protein gives an information about the excess number of cosolvent molecules inside the region defined by \( r \). Water molecules being smaller than cosolutes have a higher presence in the vicinity of protein, which is apparent from the negative dip in preferential binding parameter between 1-2 Å. This region corresponds to the peaks in radial distributions of water shown in Figure 5.1. As one moves away from the protein, the bigger urea molecules are no longer excluded and we can see that for urea the preferential binding parameter ramps up to a positive value of 7 at 2.5Å and stays there until 10 Å, indicating a preferential binding of urea. The slight changes in the value of \( \Gamma_{23}(r) \) after 2.5Å is attributed to second co-ordination shell of urea which extends until 6 Å.
Figure 5.2 $\Gamma_{23}$ plotted as a function of $r$, the distance from protein surface separating local from bulk domain. The graphs are average values obtained for first 2 ns for following systems: A. RNase T1 in water and urea B. RNase T1 in water and glycerol C. RNase T1 in water and arginine.

A close inspection of these curves also reveals a third co-ordination shell for urea. However, the error in preferential binding parameter if we do not take these into account is ±0.3 which is smaller than statistical error in our measurements ±1.0 the measurement of which will be outlined in following sections. The $\Gamma_{23}(r)$ value for glycerol and arginine also have an initial negative dip followed by an increase to a higher value around 4 Å. However, in both cases the values remain negative indicating preferential binding to water. In all the three cases the values become a constant after 6 Å and this distance can be taken as the location of boundary separating local and bulk domain for these three additives. It should be noted that the local domain increases with size of molecule. However, definition of local/bulk separating distance at a higher distance than actual does not change the value of the preferential binding parameter.

Table 5.2 shows the results of $\Gamma_{23}$ at a distance of 8Å for systems under study. The calculated average values, for first two nanoseconds and for the entire run obtained from MD simulations, are compared with the experimental values obtained from literature and from the experiments conducted in Trout lab by Curt Schneider. Experimental
measurements for RNaseT1 are available for only one additive: urea. RNase T1 being a mutant is expensive and therefore experiments were not performed on this protein. Experimental results for Lysozyme for urea are available and have been listed here along with their statistical error. It should be noted that experimental results in this case are inconsistent and therefore only provide a guide line for validating our approach. The experimental measurement technique used for measurement in the literature was dialysis or densitometry, while experiments conducted in Trout lab used vapor pressure osmometry. The discrepancy in the measurements could be attributed to the method of measurement.

Table 5.2 Table showing Preferential Binding Parameter obtained from 1. MD simulations average for first 2 ns, 2. MD simulations average for entire run, 3. Experimental values extrapolated to the concentration of interest (34) (35) (36) (37). (Values marked * were obtained from experiments and analysis performed by Curt Schneider in Trout Lab using Vapor Pressure Osmometry)

<table>
<thead>
<tr>
<th>System</th>
<th>Cosolvent</th>
<th>Length of run (nanoseconds)</th>
<th>$\Gamma_{23}$ for first 2 nanoseconds</th>
<th>$\Gamma_{23}$ for entire run</th>
<th>$\Gamma_{23}$ experimental value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNaseT1</td>
<td>Urea</td>
<td>19</td>
<td>7.7±2.9</td>
<td>13.8±0.9</td>
<td>6.4 (34)</td>
</tr>
<tr>
<td></td>
<td>Glycerol</td>
<td>15</td>
<td>-0.3±1.7</td>
<td>1.2±0.6</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Arg+</td>
<td>10</td>
<td>-3.4±4.8</td>
<td>2.5±2.5</td>
<td>NA</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>Urea</td>
<td>10</td>
<td>7.3±2.0</td>
<td>8.6±0.8</td>
<td>6.3±1.0 (35), 1.45±0.2*</td>
</tr>
<tr>
<td></td>
<td>Glycerol</td>
<td>10</td>
<td>-0.5±2.2</td>
<td>2.1±1.0</td>
<td>-1.6 (36), -6.17±0.5*</td>
</tr>
<tr>
<td></td>
<td>Arg+</td>
<td>4</td>
<td>-0.85±3.4</td>
<td>1.4±3.0</td>
<td>2.3×10^3 (37)</td>
</tr>
<tr>
<td>RNase T1 (constrained)</td>
<td>Urea</td>
<td>7</td>
<td>9.4±1.4</td>
<td>8.0±0.8</td>
<td>6.4 (34)</td>
</tr>
</tbody>
</table>

The results show that values obtained from MD simulations agree with experimental values within experimental error for first 2 nanoseconds. For example, in case of the system containing protein lysozyme with additive urea, calculated value of 7.3±2.0 matches very well with experimental value of 6.3 (35). Previous work on the preferential binding parameter has confirmed similar observations for RNase T1 with urea and RNase A with glycerol (4). This agreement with experimental values for this wide range of systems for negative as well as positive values, establishes the validity of the method for short runs.
However, for extended runs the calculated values differ from experimental values. The values are mostly higher as compared to the experimental values and in order to establish the validity of this method it is essential to study the later part (after 2 nanoseconds) of simulations to understand why there are deviations from observed experimental values. For Lysozyme, the calculated values agree with experimental values to a limited extent and do not show large deviations as in case of RNase T1 with urea.

The block averages for preferential binding parameters with blocks of 200 picoseconds are shown for RNase T1 in urea, glycerol and arginine in Figure 5.3, Figure 5.4 and Figure 5.5 respectively. The block average values show large fluctuations, with the deviations going beyond the absolute values for $\Gamma_{23}$ for the last two systems. The instantaneous values, not reported here, show even larger deviations. For the system of RNase T1 with urea the fluctuations in block average (averaging over blocks of 200ps) range from 2 to 25, while the cumulative average (taking cumulative average from the beginning of the equilibrated state up to the given time) value reaches a constant value of 13.6. The block average values for this system with blocks of 5 ns give a value of $13\pm0.5$. This value is significantly higher than the value of 7.2 obtained in the first two nanoseconds. Looking at the plot of cumulative average we can conclude that the system has reached equilibrium and gives an average value that is almost double that of the experimental $\Gamma_{23}$.

The block average $\Gamma_{23}$ for glycerol ranges from -12 to +11 while the same for Arg$^+$ ranges from -13 to +21. These two systems show results that are different from the system with urea as cosolvent. The cumulative average in this case does not go to a constant value. In case of Arg$^+$ the average value is not only higher for entire run but also changes sign from negative to positive. These systems are clearly not equilibrated and longer runs are needed to get better averages in this case. It should be noted that the simulation time for these systems was less (15 and 10 ns) as compared to the system with urea.
Figure 5.3 Time variation in the preferential binding coefficient with block averages over 200 nanoseconds for system RNase T1 with water and urea.

Figure 5.4 Time variation in the preferential binding coefficient with block averages over 200 nanoseconds for system RNase T1 with water and glycerol.

Figure 5.5 Time variation in the preferential binding coefficient with block averages over 200 nanoseconds for systems RNase T1 with water and arginine.
The time variation for lysozyme and two additives urea and glycerol is plotted in Figure 5.6 and Figure 5.7. Comparison of RNase T1 and lysozyme for both additives from above figures and Table 5.2 shows that, preferential binding parameter values for lysozyme for extended run do not differ greatly from those obtained in the first 2 nanoseconds. RMSD analysis elaborated in section 5.5 dwells on the structural stability of lysozyme and helps identify the cause of fluctuations to some extent.
The above plots illustrate the importance of having a large number of solvent and protein configurations in order to sample the entire ensemble space so as to obtain better ensemble averages. The protein in the above figures was in its native state and being unconstrained was able to sample all its configurations. The structural fluctuations in the native state of proteins have been observed on a much larger scale of 1 microsecond and these simulations highlight the importance of protein dynamics along with solvent dynamics. (36) The effect of solvent dynamics is clear from the fact that system with small additive urea as cosolvent has sampled all conformations within 5 ns while the other two systems have not. Glycerol and Arg+ being large molecules are expected to have lower diffusivity and therefore it takes longer time for the system to reach equilibrium.

![Graph](image)

Figure 5.8 Time variation in the preferential binding coefficient with block averages over 200 nanoseconds for system with RNase T1 with water and urea, with the protein harmonically constrained.

In order study further the effects of protein dynamics on the system, another system was studied with same protein RNase T1 and cosolvent urea, but with entire protein constrained to its minimized structure with a force of 10 Kcal/mol.Å². As listed in Table 5.2 the average preferential parameter for this run is 9.4 for first 2 nanoseconds and 8.0 for entire run. The simulation time for this run was restricted to 6 nanoseconds considering that the system is equilibrated with respect to $\Gamma_{23}$. Figure 5.8 shows the block average and cumulative average for this constrained protein trajectory. The cumulative average reaches a constant value of ~8 within 1 nanosecond and is within ±0.5 of that
value for further times. Thus, constraining the protein greatly reduces the fluctuations in the measured preferential binding parameter. The resulting average value of 8.0 does not show a great deviation from the observed experimental value of 6.4. The small deviation can be explained by the fact that constraining the entire protein restricts side chain motion and prevents any folding/unfolding dynamics which might lead to change in surface accessible area of side chains and therefore binding.

5.3 Local Preferential Binding Parameter

Constituent group preferential binding parameters were calculated using the algorithm in section 4.6 for the system of RNase T1 and urea, with averaging over first 2 nanoseconds and then the entire run. Figure 5.9 shows these results in terms of number of water and urea molecules co-ordinated with each contributing group. The contributing groups in this case are the amino acids and the backbone. The figure shows the average number of molecules co-ordinated for each type of group. For example, the data point for Serine is average of six Serine residues that are part of RNase T1. The solid line in the figure represents the bulk concentration i.e. any point on this concentration line will have same ratio of urea and water molecules as the bulk solvent. A residue located above this line has higher co-ordination for urea as compared to bulk and therefore shows a higher preferential binding for urea. Similarly, a residue located below the concentration line shows a lower preferential binding for urea. The protein backbone makes a separate category and lies above the concentration line. (In order to fit the high co-ordination number of protein backbone in the graph, each of the co-ordination number was divided by 10.)

There is no major difference in the nature of the preferential binding for the contributing groups for first 2 nanoseconds and for the entire run. This is evident from the fact that groups do not cross the concentration line, although in few cases the groups go towards or away from concentration line. Amino acid residues with polar side chains like Lys, Glu, Gin, Ser, Asp and Asn show preferential binding for water, while non-polar ones like Phe, Leu, Gly, Val, Ile and Cys show a preferential binding for urea. Further, some
residues like Trp and Ala, which lie away from concentration line, show a stronger tendency to preferentially bind to water and urea respectively, while others show only a weaker tendency and lie close to concentration line.

The location of a residue at a distance further away from origin shows that it co-ordinates a higher number of molecules and is exposed to the solvent. For the first two nanoseconds there are many groups lying closer to origin and slightly on the side showing preferential binding for urea. For the entire run some of these groups, namely Arg, His, Leu and Phe, shift away from origin i.e. have a higher co-ordination, showing exposure to solvent. The presence of urea as a cosolvent results in exposure of these groups and therefore changes in protein structure. It was observed that as a residue lies further away from the origin there is more tendency to preferentially bind to water molecule. Thus, reinforcing the fact that protein folds in a way that hydrophilic groups are exposed and surface area of hydrophilic group is increased.

Figure 5.10 shows the plot of contributing group preferential binding for Serine residues, which have been numbered according to their location in the protein sequence. The top graph shows results for first 2 nanoseconds and the bottom graph shows results for entire run. Corresponding error parts are also shown, showing large uncertainty in results for first 2 nanoseconds as compared to entire run. For first 2 nanoseconds, Serine residues are present on both sides of concentration line, but those that lie above are closer to concentration line and error bars touch the concentration line. For entire run, all serine residues lie either on or below the concentration line, showing neutral or preferential water binding and resulting in Serine lying considerably below the concentration line in Figure 5.9. This behavior can be explained on the basis that the Serine side chain consists of aliphatic hydroxyl group and is uncharged polar and capable of forming hydrogen bond. Moreover, the nature of binding behavior does not agree for sequential neighbors: for bottom graph in Figure 5.9, Ser12 and Ser14 lie on concentration line while Serine 13 lies below the line, implying that binding behavior is highly local. Table 5.3 lists all the Serine residues with their sequence number along with number of water and urea molecules co-ordinated. Serine residues that lie below concentration line are highlighted.
Figure 5.9 Local preferential binding parameter for system: RNase T1 in urea and water; top figure represents averages for first 2ns and the bottom figure represents averages for entire run. The number of water molecules co-ordinated is plotted against the number of urea molecules co-ordinated for each contributing group. The dark line is the concentration line for the bulk. The location of a group above the concentration line shows a preferential binding to urea as compared to water and vice-versa. For the backbone showed by symbol B, the number of urea and water molecules were divided by ten to keep the axes in lower range so as to preserve details for amino acid groups.
Figure 5.10 Local preferential binding parameter of Serine residues for system: RNase T1 in urea and water. The top figure shows averages for first 2 ns and the bottom figure shows averages for entire run. The number of water molecules co-ordinated is plotted against the number of urea molecules co-ordinated for residue Serine, along with error bars. The dark line is the concentration line for the bulk. Serine residues have been labeled according to their number in protein sequence.
Table 5.3 Number of water and urea molecules co-ordinated for each serine residue identified by its number in sequence. The numbers represent averaging over the entire run.

<table>
<thead>
<tr>
<th>Residue Name</th>
<th>Residue Number</th>
<th>Water Molecules Coordinated</th>
<th>Urea Molecules Coordinated</th>
<th>Accuracy for Water molecules (%)</th>
<th>Accuracy for Urea molecules (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SER</td>
<td>8</td>
<td>5.386</td>
<td>0.295</td>
<td>13.0</td>
<td>13.0</td>
</tr>
<tr>
<td>SER</td>
<td>12</td>
<td>15.735</td>
<td>0.327</td>
<td>11.8</td>
<td>21.4</td>
</tr>
<tr>
<td>SER</td>
<td>15</td>
<td>12.912</td>
<td>1.084</td>
<td>7.4</td>
<td>11.1</td>
</tr>
<tr>
<td>SER</td>
<td>14</td>
<td>126.121</td>
<td>2.483</td>
<td>7.5</td>
<td>8.7</td>
</tr>
<tr>
<td>SER</td>
<td>17</td>
<td>30.099</td>
<td>0.614</td>
<td>10.7</td>
<td>15.2</td>
</tr>
<tr>
<td>SER</td>
<td>37</td>
<td>41.238</td>
<td>0.831</td>
<td>22.9</td>
<td>17.0</td>
</tr>
<tr>
<td>SER</td>
<td>31</td>
<td>10.233</td>
<td>0.274</td>
<td>9.3</td>
<td>6.2</td>
</tr>
<tr>
<td>SER</td>
<td>33</td>
<td>102.792</td>
<td>1.574</td>
<td>5.2</td>
<td>5.2</td>
</tr>
<tr>
<td>SER</td>
<td>54</td>
<td>61.224</td>
<td>1.153</td>
<td>12.1</td>
<td>12.1</td>
</tr>
<tr>
<td>SER</td>
<td>63</td>
<td>11.099</td>
<td>0.222</td>
<td>31.1</td>
<td>27.8</td>
</tr>
<tr>
<td>SER</td>
<td>64</td>
<td>56.336</td>
<td>1.123</td>
<td>9.0</td>
<td>12.4</td>
</tr>
<tr>
<td>SER</td>
<td>69</td>
<td>46.444</td>
<td>0.932</td>
<td>16.1</td>
<td>14.4</td>
</tr>
<tr>
<td>SER</td>
<td>72</td>
<td>27.872</td>
<td>0.506</td>
<td>19.0</td>
<td>18.0</td>
</tr>
<tr>
<td>SER</td>
<td>96</td>
<td>39.564</td>
<td>0.973</td>
<td>7.4</td>
<td>7.4</td>
</tr>
</tbody>
</table>

Figure 5.11 Local binding behavior of urea and water with the amino acid backbone and side chains in RNase T1: obtained from previous simulations (4). The labels are the one-letter code for each amino acid side chain, and "B" is the protein backbone. The line denotes the bulk urea concentration. In addition to the protein backbone and Ser, the hydrophobic amino acids Cys, Gly, Leu, Phe, Pro, Tyr, and Val all preferentially bind urea, while the hydrophilic Asp preferentially binds water.
Lastly, we compare the results obtained for first 2 nanoseconds from Figure 5.9 with the results from previous simulations obtained for 2 nanosecond runs, shown in Figure 5.11 (4). The backbone lies above the concentration line in both cases and so do the hydrophobic amino acids Gly, Leu, Phe, Pro and Val, while hydrophobic Asp lies below. A big difference is observed for Serine group which lays preferentially binds to water in Figure 5.9 while it binds urea in Figure 5.11. Similarly, small differences are also observed for residues Glu, Gln, Arg and Cys.

Figure 5.12 shows the local preferential binding parameter for system of RNase T1 and glycerol. Backbone in this case lies below the concentration line, showing preferential binding for water. Similar to system containing urea, Lys shows a very high co-ordination number with preference for water and Asp & Glu show a strong hydrophilic behavior. A big difference in co-ordination number for the top (2 nanosecond) and bottom (entire run) is observed for Gln and Phe. Gln reduces its co-ordination and becomes slightly hydrophilic when averaging is done for entire run. Phe increases its co-ordination number to a great extend but does not cross concentration line. The movement of phenol in this graph can be attributed to favorable changes in conformations that increase the surface accessible area for Phe. A study of individual Gln and Phe residues reveals that the averages are affected by only one residue in each case namely Gln85 and Phe48.
Figure 5.12 Local preferential binding parameter for system: RNase T1 in glycerol and water; top figure represents averages for first 2ns and the bottom figure represents averages for entire run. The number of water molecules co-ordinated is plotted against the number of glycerol molecules co-ordinated for each contributing group. The dark line is the concentration line for the bulk. The location of a group above the concentration line shows a preferential binding to glycerol as compared to water and vice-a-versa. For the backbone showed by symbol B, the number of glycerol and water molecules were divided by ten to keep the axes in lower range so as to preserve details for amino acid groups.
5.4 Error Analysis

The error analysis based on discussion in section 4.4 is presented for the system of RNase T1 with urea. The preferential binding parameter for protein is analyzed for error by calculating ratio \( p(\tau_b) \) and plotting it against \( \tau_b^{1/2} \). Figure 5.13 shows that as \( \tau_b \) goes to infinity, the value of \( p(\tau_b) \) goes to a constant value which is around 250 picoseconds. This means that only about one configuration in 250 picoseconds run contributes completely new information to the average. The total run is 19 nanoseconds for this system and the RMS fluctuation is 8.1 giving the accuracy of:

\[
(\frac{250}{19000})^{1/2} \times 8.1 \equiv 0.9
\]  

(5.1)

This is the error in the values reported in table 5.2. The numbers obtained for other systems were of same order for the preferential binding parameter. A confidence level of \( \pm 0.9 \) represents a confidence limit in free energy of about 0.6 Kcal/mol for an additive following equation 3.5, and this is typical of molecular simulations. Analysis similar to above when applied to the local preferential binding coefficient gives the number of correlated steps to be around 600 as shown in Figure 5.14. The error bars plotted according to this analysis are shown in Figure 5.10 and listed in Table 5.3.

For the case of RNase T1 and urea, a constrained system showed a statistical inefficiency of 80 which is about one-third times that reported for unconstrained run. This brings down the simulation time by a factor of three, which is worthwhile considering the time taken for computer simulation of one system is about a month. The system containing lysozyme in urea shows statistical inefficiency \( \sim 63 \) which is much lower than \( \sim 250 \) calculated for RNase T1, resulting in higher accuracy for the preferential binding parameter results shown in Table 5.2.
Figure 5.13 Plot of ratio $p(\tau_s) = \frac{\tau_s \sigma^2(<A_s>)}{\sigma^2(A)}$ against $\tau_b^{1/2}$ for MD simulation run of RNase T1 in urea. In this plot the quantity A is preferential binding parameter. The statistical efficiency s is the value at which the plot goes to a constant plateau which is shown by a dotted line. The dotted line shows the value where a plateau is approached.

Figure 5.14 Plot of ratio $p(\tau_s) = \frac{\tau_s \sigma^2(<A_s>)}{\sigma^2(A)}$ against $\tau_b^{1/2}$ for MD simulation run of RNase T1 in urea. In this plot the quantity A is local preferential binding parameter for backbone. The dotted line shows the value where a plateau is approached.
5.15 Plot of ratio $p(\tau_b) = \frac{\tau_b \sigma^2(\langle A_b \rangle)}{\sigma^2(A)}$ against $\tau_b^{1/2}$ for MD simulation run of lysozyme in urea. In this plot the quantity $A$ is the local preferential binding parameter for protein. The dotted line shows the value where a plateau is approached.

### 5.5 RMSD Analysis

As shown in section 5.2, the values of preferential binding coefficient show very high fluctuations and can be attributed to conformation changes in protein structure. The root mean square deviation (RMSD) averages the deviation for all atoms in protein. Plot of RMSD against time in nanoseconds is shown in Figure 5.16 and Figure 5.17. During the simulations, the protein was not fixed in any way and was free to rotate or translate through the box, therefore for calculating RMSD it is necessary to re-center and reorient this protein so that RMSD captures only structural changes like partial unfolding and not changes due to other motions.
The protein RNase T1 has a very high RMSD for all three additives: the highest value for arginine is 6.6, for urea 6.2 and for glycerol 5.8. The RMSD values for RNase T1 without any cosolvent are smaller than those with cosolvent, reaching a value of 4 at 7 nanoseconds up to which the results are available. The presence of cosolvent has an effect on the protein conformation resulting in unfolding to some extent. The time scales for our simulations are very small for such transformation to take place completely. Only partial conformational changes are observed through RMSD. On the other hand RMSD values for lysozyme are very low: below 3 for cosolvent urea and arginine and ~3.5 for cosolvent glycerol. Comparison between the two proteins show that RNase T1 has higher structural instability, and the presence of cosolvents show structural changes that are uncommon at time scale of simulation.

The fluctuation in protein conformations are correlated with the fluctuations observed in the preferential binding parameter. For RNase T1 large fluctuations in protein conformations results in large number of protein-cosolvent conformations which cannot be a sampled even in a long run of 10 nanoseconds. On the other hand lysozyme shows a highly stable structure and limited fluctuations, resulting in highly stable values for preferential binding parameter. These findings are consistent with very low accuracy obtained for the preferential binding parameter results for RNase T1 when compared to lysozyme with same additive and same duration of simulation.
Figure 5.16 The RMSD (in units of Å) results for RNaseT1 in water and solvents: urea, glycerol and arginine. The fourth system is protein in water without cosolvent. The protein was recentered and reoriented for these calculations.

Figure 5.17 The RMSD (in units of Å) results for Lysozyme in water and solvents: urea, glycerol and arginine. The protein was recentered and reoriented for these calculations.
6 Conclusion and Future Work

A quantitative method based on single trajectory MD simulations, with all atom potential models was tested and validated with limitations for two proteins. MD simulations were run over extended time of about 10 nanoseconds, which revealed some properties of the system not seen in a short run. The results of simulations match experimental data for the initial trajectory of 2 nanoseconds after which we observe deviations. A constrained protein trajectory has successfully overcome this problem with positive agreement with experimental values for extended runs and better accuracy.

Protein dynamics as well as solvent dynamics play an important role in determining $\Gamma_{23}$. In order to sample all conformations in a system with a protein showing structural instability longer runs are needed. The radial distribution functions, which are an intermediate step to calculation of preferential binding parameter, reveal important features of solvent structure around protein that can be used to develop mechanistic models of binding. The presence of co-ordination shells helps us identify the bulk and local region.

Local preferential binding results show that the contributing groups within the protein show a variety of preferential binding. The preferential binding not only depends on the chemistry of the contributing group but also its location within the protein. Thus the structure of protein at levels: primary, secondary, tertiary plays a role in binding behavior of contributing groups. Moreover, preferential binding is a highly local phenomenon: contributing groups, which are sequential neighbors and have same chemistry, do not show similar binding nature. Therefore, any group preferential binding theory should be based on the structure as well as the sequence of the protein. A comparison of the local preferential binding for first 2 nanoseconds with the entire run reveals a few changes in preferential binding as well as total co-ordination number during the run that imply changes in protein conformation. Some groups are exposed while some are hidden due to contact with a mixed solvent.
The above work helps us understand the effects of additives on protein and at the same time raises several issues which should form basis of future work. MD simulations used standard CHARMM potentials and the results are highly dependent on these potentials. One strategy to get better prediction of preferential binding using simulations is to fine tune the potential parameters for additives of interest for our model system and test it on other systems.

There is a clear need for longer simulations in order to sample all conformations of protein as well as solvent. With the current progress in parallel computing and development of simulation methods, it will be possible to have sufficiently long trajectories and have better averages in the future.

Constraining was proved to be an effective way to get equilibrium values in shorter runs. However, the system in that case does not mimic real proteins. In order to test the effect of constraining on the equilibrium value of the preferential binding parameter, protein can be constrained partially by constraining only the backbone and allowing free motion of side chains.

Lastly, there is a need for development of group preferential binding theory based on knowledge of contributing groups and the structure of the protein. A model based on calculations similar to above that will be able to predict the protein stabilization without experimental screenings holds the key to rational additive design to prevent aggregation of proteins. Moreover, a technique that sufficiently reduces time required in sampling position space of the protein and additives will result in analysis of larger proteins.
Bibliography

21. Light Scattering arising from Composition Fluctuations in multi-component systems.  
22. Charmm: a program for macromolecular energy, minimization, and dynamics calculations.  
23. Comparison of simple potential systems for simulating liquid water. W. L. Jorgensen,  
Vol. 79, pp. 926-935.
27. Variability in the pKa of histidine side-chains correlates with burial within proteins.  
29. Free Energy Calculations: Applications to chemical and biological phenomenon.  