Molecular Modeling of Solute/Co-Solvent/Water Preferential Interactions: Toward Understanding the Role of Hydration and Co-solvent in Weak Protein-Protein Interactions

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of The Ohio State University

By
Hamsa Priya Mohana Sundaram, B.Tech., M. E.
Graduate Program in Chemical and Biomolecular Engineering

The Ohio State University

2011

Dissertation Committee:
Prof. Michael E. Paulaitis, Advisor
Prof. Aravind Asthagiri
Prof. Dilip Asthagiri
Prof. Sherwin J. Singer
ABSTRACT

For bulk production of high purity protein drugs to meet FDA standards, pharmaceutical industries typically process proteins by repeated precipitations and crystallizations. The solution conditions for protein phase separation are routinely determined by trial and error. A thermodynamic model of underlying weak protein-protein interactions will help in a rapid screening of solution conditions in obtaining protein crystals of desired quality and stability. The principal objective of the thesis is, therefore, to understand the fundamental molecular level interactions among different components of protein solutions: protein, water and co-solvents, in particular, the role of hydration and co-solvent preferential interactions on protein-protein interactions.

Spatial heterogeneities in protein chemistry and surface topography results in uneven specific hydration of protein surface, which alters the protein-protein interactions by eliminating some complimentary configurations. The dynamics of water at such specific hydration sites was examined in terms of average water residence times and average vacancy times and found to have little impact on protein-protein interactions. The influence of local heterogeneities in surface charge and surface roughness on specific hydration and water dynamics have also been examined. A detailed investigation of the effect of surface curvature on hydration revealed that hydration of a concave
surface is thermodynamically expensive than the hydration of a chemically equivalent convex surface. The concave surface is found to remain hydrated only when the interaction between the water and constituent surface atoms are attractive.

Protein phase separation is typically induced by adding a precipitating agent or co-solvent, such as inorganic salts or organic compounds: alcohols, polyols, or polyethylene glycol (PEG). Addition of a cosolvent to protein solutions alter the preferential hydration of proteins depending upon its affinity to interact with proteins. Most of the salting-in agents, preferentially partition onto the protein surface, while the salting-out agents preferentially exclude from the surface.

The molecular preferential interactions of organic co-solvents like methanol, ethanol, glycerol and urea with molecular solutes (methane, ethane, neopentane, tetra-methyl ammonium ion) and a model globular protein (lysozyme) have been evaluated using statistical thermodynamic solution theories like Kirkwood-Buff theory and quasi-chemical theory. The characteristic convergence problem in the former due to long range oscillations in the radial distribution functions of water and co-solvent around a solute has been obviated by smearing the distribution assuming multiple reference centers around each heavy atom of the solvent molecules. Quasi-chemical theory which distinguishes local and non-local interactions naturally provides the appropriate molecular framework for modeling local preferential hydration and co-solvent preferential interactions at specific sites on protein surface. From the light scattering measurements on interactions between lysozyme and PEG with hydroxyl (hPEG) and methyl (mPEG) end groups, it is found that increasing the end-group hydrophobicity of PEG promotes the preferential interaction of PEG with lysozyme and thereby, stabilizes the dissolution of lysozyme in the solution.
To my beloved parents
ACKNOWLEDGMENTS

First and foremost I express my sincere thanks to my advisor Prof. Michael Paulaitis for his constant support and guidance over these years. I have always been awed by his astute interpretation and judgement on the accuracy of the data. I understand the key to such good clarity in thinking is what he always emphasizes to me ”Think Physically, Not Mathematically”. He has been very considerate and patient towards me, while I have been taking baby steps to meet his standards, which I hope to meet some day. I enjoyed being a teaching assistant for his classes, they were great opportunities to acquire physical understanding of thermodynamics.

I want to thank Prof. Dilip Asthagiri for in-numerous helpful discussions, especially, in the initial years. I learnt a lot during the short period I worked in his lab at JHU. I extend my thanks to Dr. Jindal Shah, Purusha and Safir for many enjoyable scientific and nonscientific arguments and making my stay at JHU most memorable.

I would like to thank Prof. Sherwin Singer and Prof. Aravind Asthagiri for agreeing to be on my thesis committee. I owe special thanks to Prof. Hank Ashbaugh for several helpful discussions on smearing work. I appreciate the help provided by the personnel at Ohio Supercomputer Center, in particular to Jim Giuliani, Doug Johnson and Yuan Zhang, in solving computational issues. I thank my past and current labmates Arturas, Sowmi, Shikha, Ghalib, Chaofang, Nicole, Kitty, Amy for all good times.
I acknowledge that my Master’s thesis advisor, Prof. Giridhar Madras, and undergraduate teachers notably Prof. K. S. N. Raju and Srikanth, were the main advocators to pursue PhD. My special thanks to my best friends Siva, Uday, Satish, Nithya, Lavanya, Dilip, Gautham, Gopi, and Suresh for motivating and boosting my morale. I thank my close friends Priya, Preshit, Jayashri and Hitesh, and also the entire Desi gang in the department for making my stay at OSU enjoyable.

During this period, I also had the opportunity to learn guitar, my long time wish. I thank Tom for his patience in teaching me music, and Ramya for gifting me a guitar, one of the most cherished gift. I appreciate the love and motivating support of Padma and Kumar.

This thesis is a testimony of love and confidence my parents, Mohana Sundaram and Ambika, have on me. My Dad has always been a greatest inspiration to me, I am very proud to write that he is the first reader of my thesis. Without my Mom’s support the thesis would have been only a dream. Understanding me in and out, she gave me the required emotional support, helping me prioritize things and work effectively, during my difficult times. I dedicate this thesis to my parents for their unconditional love and affection.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
<tr>
<td>Chapters:</td>
<td></td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2. Distinguishing Thermodynamic and Kinetic Views of the Preferential Hydration of Protein Surfaces</td>
<td>7</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>7</td>
</tr>
<tr>
<td>2.2 Theory</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Methods</td>
<td>13</td>
</tr>
<tr>
<td>2.4 Results and Discussion</td>
<td>16</td>
</tr>
<tr>
<td>2.5 Conclusions</td>
<td>24</td>
</tr>
<tr>
<td>3. A Simulation Study on the Wetting Characteristics of a Sinusoidal Hydrophobic Wall</td>
<td>27</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>27</td>
</tr>
<tr>
<td>3.2 Methods</td>
<td>31</td>
</tr>
<tr>
<td>3.2.1 Molecular Simulations</td>
<td>31</td>
</tr>
<tr>
<td>3.2.2 Analysis</td>
<td>33</td>
</tr>
<tr>
<td>Chapter</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>8.</td>
<td>Effect of PEG End-Group Hydrophobicity on Lysozyme-Lysozyme Interactions</td>
</tr>
<tr>
<td>8.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>8.2</td>
<td>Theory</td>
</tr>
<tr>
<td>8.3</td>
<td>Materials and Methods</td>
</tr>
<tr>
<td>8.4</td>
<td>Results and Discussion</td>
</tr>
<tr>
<td>8.5</td>
<td>Conclusions</td>
</tr>
<tr>
<td>9.</td>
<td>Static Light Scattering Measurements of Protein Osmotic Cross Virial Coefficients</td>
</tr>
<tr>
<td>9.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>9.2</td>
<td>Theory</td>
</tr>
<tr>
<td>9.3</td>
<td>Materials and Methods</td>
</tr>
<tr>
<td>9.4</td>
<td>Results and Discussion</td>
</tr>
<tr>
<td>9.5</td>
<td>Conclusions</td>
</tr>
<tr>
<td>10.</td>
<td>Conclusions and Future work</td>
</tr>
<tr>
<td>10.1</td>
<td>Conclusions</td>
</tr>
<tr>
<td>10.2</td>
<td>Future work</td>
</tr>
</tbody>
</table>

Appendices:

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.</td>
<td>Supplementary Data for Preferential Interactions of Molecular Solutes and the Proximal Group Contributions</td>
<td>167</td>
</tr>
<tr>
<td>B.</td>
<td>Supplementary Data for Quasi-Chemical Theory of Preferential Interactions in Aqueous Solutions</td>
<td>173-178</td>
</tr>
<tr>
<td>B.1</td>
<td>Error estimation of gaussian approximation for the outer-shell contribution to the solute excess chemical potential</td>
<td>173</td>
</tr>
<tr>
<td>C.</td>
<td>Supplementary Data for Site Specific Co-solvent Preferential Interactions on Protein Surface through Quasi-Chemical Theory</td>
<td>178</td>
</tr>
</tbody>
</table>

Bibliography | 187 |
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Interaction part of second virial coefficient, $\beta_{22}$, as a function of hydration conditions for lysozyme at pH 7 and ionic strength 0.007 moles/liter and staphylococcal nuclease at pH 6.5 and ionic strength 0.01 moles/liter</td>
<td>23</td>
</tr>
<tr>
<td>3.1 Simulation box length in the z-direction (in Å) for simulations at different LJ pair potentials defined by the scaling parameter $\lambda$ (Eq. 3.1).</td>
<td>33</td>
</tr>
<tr>
<td>4.1 Classification of constituent groups of the amino acids with their solvent excluded volume radii</td>
<td>55</td>
</tr>
<tr>
<td>4.2 PMV of the N-acetyl amino acid amides at infinite dilution in water at 300 K, determined from simulations ($V^\infty_{\text{Sim}}$) and group contributions ($V^\infty_{\text{GC}}$) using KB theory. PMV contributions from side-chain ($V^\infty_{\text{SC}}$) and backbone ($V^\infty_{\text{B}}$), including Cα carbon atom, of amino acids are also listed.</td>
<td>57</td>
</tr>
<tr>
<td>5.1 Number of co-solvent and water molecules in the simulation as a function of the co-solvent concentration</td>
<td>72</td>
</tr>
<tr>
<td>5.2 Preferential interaction parameters for methanol (co-solvent) with solutes, neopentane and TMA, in aqueous methanol solutions, evaluated as the sum of preferential interaction parameters for the constituent groups of the solutes. Standard deviations obtained by dividing the simulation trajectory into five equal parts and block averaging.</td>
<td>76</td>
</tr>
<tr>
<td>5.3 Preferential interaction parameters for different co-solvents with solutes, neopentane and TMA in 20 mol % aqueous co-solvent solutions, evaluated by smearing the co-solvent effective grand canonical distribution around the constituent groups of each co-solvent.</td>
<td>79</td>
</tr>
</tbody>
</table>
6.1 System Specifications

6.2 QC contributions to $\mu^{ex}$ (Eq. 6.13) for methane in aqueous methanol solutions as a function of methanol mole fraction computed using mean binding energies and fluctuation contributions obtained from gaussian descriptions of the conditional probability distributions in Figs. 6.3 and B.3 with inner-shell radii: $\lambda_w = 3.3$, $\lambda_m = 3.6$, and $\lambda_h = 3.9$ Å. All free energies are in kcal/mol.

6.3 Methane-methanol preferential interaction parameter $\Gamma_{sc}$ (Eq. 6.15) in aqueous methanol solutions as a function of methanol mole fraction, calculated from the slope of $\mu^{ex}(\rho_c)$ in Fig. 6.5, compared with $\Gamma_{sc}$ obtained from KB theory (Eq. 6.2) using the methane-water and methane-methanol rdfs with a cutoff radius of 6.0 Å. Units for $\Gamma_{sc}$ are number of methanol molecules/methane molecule.

6.4 QC contributions to the methane-methanol preferential interaction parameter $\Gamma_{sc}$ for 10 mol% methanol/water solution as a function of the inner-shell radius for water, $\lambda_w$. The corresponding radii for the methyl and hydroxyl groups of methanol are $\lambda_m = \lambda_w + 0.3$ Å, and $\lambda_h = \lambda_w + 0.6$ Å, respectively. Units for $\Gamma_{sc}$ are number of methanol molecules/methane molecule.

7.1 Protein sites of interest for local co-solvent preferential interaction analysis.

7.2 Classification of molecular groups on protein surface with their water excluded-volume radii, identified as the smallest radial distance at which the proximal radial distribution of water around these groups reach unity.

7.3 $R_i$ for the heavy atoms of urea and methanol, defined as one half of the distance at which radial distribution function $g_{ii}(r)$ in bulk co-solvent solution first reaches unity.

8.1 The osmotic virial coefficient, $B_{11}$, osmotic cross virial coefficients, $B_{12}$ and $C_{112}$, and the polymer concentration dependence of the effective osmotic virial coefficient, $B_{eff11}$, for aqueous solutions of lysozyme (1) and hydroxy-terminated PEG (hPEG, 2) or methyl-terminated PEG (mPEG, 2) at ionic strengths of 0.01 and 0.8 M and $T = 25^\circ$C.
9.1 Comparison of osmotic virial coefficients corresponding to self and cross interactions between lysozyme, α-chymotrypsinogen A measured by static light scattering (SLS) to the literature values.

A.1 Partial molar volumes of molecular solutes(s) in neat water(i) and neat methanol(i) determined by KB approach: \( V_s^\infty = kT\kappa_0 - G_{si} \). The calculated values are compared against the values empirically predicted via group contribution technique based on the experimental data for a range of hydrophobic solutes.

A.2 Preferential interaction parameter of methanol with different solutes, in the units of number of cosolvent molecules per molecule of the solute, in 50 mol % methanol solution.

B.1 Absolute values of the errors in the estimation of outer contribution by an gaussian approximation for the distribution of binding energies of methane, at different concentrations of methanol in the solution, in an empty inner shell defined by \(-\lambda_w = 3.3 \text{ Å} ; \lambda_m = 3.6 \text{ Å} ; \lambda_h = 3.9 \text{ Å} \).

C.1 Inner shell and total preferential interaction parameters of 7 protein sites obtained through QC analysis in 2 mol% urea solution and 10 mol% methanol solution, averaged over five inner shell volumes i.e., for \(\lambda_w\) ranging from 2.6 to 3.0 Å.
LIST OF FIGURES

Figure | Page
---|---
2.1 | Natural logarithm of the ratio of average occupancy times ($\nu_1 = \tau_1/\tau_{1,b}$), average cycle times ($\nu_2 = \tau_{cyc,b}/\tau_{cyc}$), and the combination of the two times ($\nu = \nu_1\nu_2$), defined in Eq. 2.4 as a function of $\eta$ (Eq. 2.1) for hydration sites within 3.5 Å of the surface of lysozyme. Only those sites corresponding to $\eta > 2.0$ are shown. These characteristic times were obtained from MD simulations of lysozyme in TIP3P water at 300 K and 1 bar. Correlation coefficients for the linear fits of the data are 0.654, 0.176 and 0.999, respectively. | 15

2.2 | Average occupancy time, $\tau_1$, and average vacancy time, $\tau_0$, for each point in a grid filling the proximal volume within 3.5 Å of the heavy atoms on the surface of lysozyme, obtained from MD simulation of this protein in TIP3P water at 300 K and 1 bar. Solid triangles: $\eta > 2.7$; open circles: $2.7 > \eta > 2.0$; gray crosses: $\eta < 2.0$. The dashed $\tau_1 = \tau_0$ line separates sites that correspond to kinetically bound waters (solid triangles) from the sites corresponding to high water occupancies ($\eta > 2.0$) (open circles), while the dash-dot line separates the high occupancy sites from all other hydration sites (gray crosses). The inset shows only those data for $\tau_1, \tau_0 \leq 2$ ps. | 17

2.3 | Average occupancy time, $\tau_1$, and average vacancy time, $\tau_0$, for each point in a grid filling the proximal volume within 3.5 Å of the heavy atoms on the surface of staphylococcal nuclease, obtained from MD simulation of this protein in TIP3P water at 300 K and 1 bar. Solid triangles: $\eta > 2.7$; open circles: $2.7 > \eta > 2.0$; gray crosses: $\eta < 2.0$. The dashed $\tau_1 = \tau_0$ line separates sites that correspond to kinetically bound waters (solid triangles) from the sites corresponding to high water occupancies ($\eta > 2.0$) (open circles), while the dash-dot line separates the high occupancy sites from all other hydration sites (gray crosses). | 18
2.4 Ten highest occupancy sites on the surfaces of spheres with the following chemical and/or topographical features: (a) smooth dipolar surface, (b) smooth random surface, (c) rough apolar surface, and (d) rough random surface. The four spheres have a radius of 10 Å, and are filled with neutral atoms in the interior. Atoms on the surface are: neutral (green), negatively charged (-0.51 e, red), or positively charged (+0.51 e, blue). The models are constructed to be electrostatically neutral by placing an atom of opposite charge below the surface for every charged atom on the surface. The gray spheres represent water oxygens at these high-occupancy sites. Lennard Jones parameters for all atoms are \( \sigma = 3.4 \) Å and \( \epsilon = -0.12 \) kcal/mol.

2.5 Average occupancy time, \( \tau_1 \), and average vacancy time, \( \tau_0 \), for the ten highest occupancy sites on the surface of four idealized (spherical) models of globular proteins obtained from MD simulations in TIP3P water at 300 K. Circles: smooth dipolar surface; diamonds: smooth random surface; crosses: rough apolar surface; triangles: rough random surface. The dashed line and the dash-dot line are defined in Figs. 2.2 and 2.3.

3.1 Sinusoidal Lennard-Jones (LJ) wall comprised of 6144 closed-packed LJ atoms separated by 2.5 Å, the center-to-center distance between adjacent atoms. The wall extends over a width of 80 Å in the x-direction, and a length of 80 Å in the y-direction. The well depth is 20 Å in the z-direction, measured from the center of the LJ atom in the bottom center of the well to the center of the LJ atom at the top. The atoms comprising the convex and the concave surfaces of the wall are highlighted in green and yellow, respectively.

3.2 Snapshots from MD simulations of the hydration of the LJ wall (Fig. 3.1) as attractive interactions between the wall atoms and water molecules are scaled from \( \lambda = 1.0 \) to \( \lambda = 0.3 \) (Eq. 3.1). Top row (left to right): \( \lambda = 1.0, 0.9, 0.8 \) and 0.7. Bottom row (left to right): \( \lambda = 0.6, 0.5, 0.4 \) and 0.3.
3.3 Top panel: Normalized water density as the function of the well height, $z$ as attractive interactions between the wall atoms and water molecules are scaled from $\lambda = 1.0$ to $\lambda = 0.3$ (Eq. 3.1). Bottom panel: Water density profile across the width of the concave well at a height of 13 Å < $z < 16$ Å for the same range of $\lambda$-scaling of the wall-water attractive interactions.

3.4 Potential of mean force of bringing a water molecule from bulk solution to a distance, $r$, from any atom in the concave (top panel) or convex (bottom panel) surfaces of the LJ wall (Fig. 3.1) as attractive interactions between the wall atoms and water molecules are scaled from $\lambda = 1.0$ to $\lambda = 0.7$ (top panel) or $\lambda = 1.0$ to $\lambda = 0.3$ (bottom panel).

3.5 Average number of water molecules in spherical observation volumes of radius 3.3 Å located 3.4 Å above the bottom center of the concave surface (solid circles) and 3.4 Å above the center of the convex surface (open circles) of the LJ wall as a function of $\lambda$ scaling of the wall-water attractive interactions. The bulk water occupancy number in an identical observation volume located 30 Å above the convex surface is shown as a horizontal dashed line. Lines connecting the points show the trends in the data.

3.6 Correlation of water occupancies in two spherical observation volumes of 3.3 Å radius located 3.4 Å above the bottom center of the concave surface (red lines) and 3.4 Å above the center of the convex surface (black lines) as a function of the center-to-center distance separating the observation volumes in the y-direction. The lines correspond to different $\lambda$ scaling of wall-water attractive interactions, with the arrows showing the direction of decreasing $\lambda$. For the red lines, $\lambda = 1.0, 0.9, 0.8$ and 0.7. For the black lines, all values of $\lambda$ between 1.0 and 0.3.

3.7 The variation of excess chemical potential of water (solid triangles) and its corresponding quasi-chemical contributions: $kT \ln x_0$ - open circles; $-kT \ln p_0$ - solid circles; outer shell contributions - open triangles – as the function of well height when water completely wets and partially dewets the wall, i.e., when the wall-water attraction is completely turned on, $\lambda = 1.0$, – top panel – and when the attraction is partially (50%) turned off, $\lambda = 0.5$ – bottom panel. The dashed line is the experimental value of excess free energy of water, -6.32 kcal/mol [1].
3.8 Logarithm of conditional probability distribution of binding energies, conditioned such that there is no water within 3.3 Å spherical observation volume about the central water oxygen, of waters in the concave well when the wall-water attraction is completely turned on, \( \lambda = 1.0 \), -- top panel -- and when the attraction is partially (50%) turned off, \( \lambda = 0.5 \) -- bottom panel -- at different well heights, \( z \): 4 < \( z \) < 5 Å - black circles; 9 < \( z \) < 10 Å - red circles; 14 < \( z \) < 15 Å - green circles; \( z = 19 < z < 20 \) Å - blue circles; \( z = 24 < z < 25 \) Å - violet circles; 29 < \( z \) < 30 Å - dark green circles; 34 < \( z \) < 35 Å - orange circles. The lines are the gaussian fit. The consecutive energy distributions are shifted by 4 units for visual clarity.

4.1 Top panel: The effective grand canonical radial distribution function for water, smeared (solid line) and unsmeared (dashed line), around N-acetyl-alanine-amide determined from the proximal radial distribution of water around its constituent groups. Bottom panel: PMV of N-acetyl-alanine-amide computed from Eq. 4.5.

4.2 Comparison of the PMV of the N-acetyl amino acid amides excluding cystine, histidine, lysine, arginine, aspartic acid and glutamic acid, calculated from simulations with their corresponding experimental values.

4.3 Comparison of PMV side-chain contributions computed from the simulation to the experimental values obtained for the zwitterionic amino acids in aqueous buffer solution.

4.4 The proximal effective grand canonical radial distribution functions for waters around proximal groups of the N-acetyl amino acid amides: methyl group (top panel), carbonyl oxygen group (bottom left panel) and nitrogen group (bottom right panel). Water proximal distributions around carbonyl oxygen in aspartic and glutamic acid and nitrogen group in lysine and arginine are distinct from others as shown in dashed lines.

4.5 Comparison of PMV side-chain contributions measured experimentally against values calculated by different approaches: 1D-RISM theory -- circles, 3D-RISM theory -- squares; simulations -- triangles. The lines are linear fits through origin (\( R^2 > 0.95 \)) with slopes of 0.54, 0.84 and 1.02 for 1D-RISM (dash-dotted line), 3D-RISM (dashed line) and simulations (solid line), respectively.
4.6 Comparison between the group contribution estimate of the PMV side-chain contributions and the experimental data for zwitterionic amino acid in aqueous buffer solution. Group contribution predictions are obtained using the group-averaged proximal correlation functions for uncharged amino acids (Eq. 4.8). . . . . . . . . . . . . . . . . . . . . . . 62

5.1 Concept of smearing the co-solvent distribution function: The location of a co-solvent molecule – in this case, urea – relative to the solute of interest is defined by a randomly selected reference point within the shaded spherical volumes centered on the different constituent groups of urea, rather than a single reference point; e.g., the center of mass. Radii defining the spherical volumes for the different constituent groups are: 2.3 Å for an oxygen atom or hydroxyl group, 2.5 Å for a nitrogen atom or amine group, and 3.0 Å for a carbon atom or a methyl or methylene group. These radii are the radial distances up to which the corresponding proximal distribution functions equal zero. . . . . . . . . . . . . . . . . . . . . . . 71

5.2 Preferential interaction parameters for methanol (co-solvent) with neopentane (solute) in the 20 mol% aqueous methanol solution as the function of cutoff radius, \( R_c \), computed using canonical radial distribution functions for methanol (center of mass) and water (solid line) and using the corresponding smeared effective grand canonical radial distribution functions (dashed line). Radii of 2.3 Å and 3.3 Å were used for smearing the water and methanol distribution functions, respectively. 73

5.3 Preferential interaction parameters for methanol (co-solvent) with the solutes, neopentane (top panel) and TMA (bottom panel), in aqueous solutions at different methanol concentrations as a function of cutoff radius, \( R_c \). \( \Gamma_{sc} \) is calculated using the smeared effective grand canonical distributions for methanol and water. Methanol concentrations (mol fraction): 0.05 - solid line; 0.10 - dotted line; 0.20 - dashed line; 0.30 - dash-dot line; 0.50 - dash-dot-dash line. . . . . . . . . . . . . . . . . . . . . . . 75
5.4 Preferential interaction parameters for different co-solvents with the solutes, neopentane (top panel) and TMA (bottom panel), in 20 mol % aqueous co-solvent solutions as a function cutoff radius, $R_c$, calculated using smeared effective grand canonical distribution functions for the co-solvents and water. Co-solvents: methanol - solid line; ethanol - dotted line; glycerol - dash-dot line; urea - dash-dot-dash line. The horizontal dashed line is the baseline corresponding to $\Gamma_{sc} = 0$. The co-solvent distributions were smeared based on their constituent groups.

6.1 Radial distribution functions for water oxygens around a methane molecule in water (solid line) and in aqueous methanol solutions at different methanol mole fractions: 0.05 (dashed line), 0.10 (dash-dot line), 0.20 (dotted line), 0.30 (long dashed line). Radial distribution functions are shifted successively upwards by 0.2 units for clarity.

6.2 Radial distribution functions for methanol methyl groups (top) and hydroxyl groups (bottom) around a methane molecule in methanol (solid line) and in aqueous methanol solutions at different methanol mole fractions: 0.05 (dashed line), 0.10 (dash-dot line), 0.20 (dotted line), 0.30 (long dashed line). Radial distribution functions are shifted successively upwards by 0.2 units for clarity.

6.3 Conditional probability distributions of binding energies for methane in water (solid circles) and in methanol (open circles) when no water oxygens are within $\lambda_w = 3.3$ Å of methane in water, and no methanol methyl and hydroxyl groups are within $\lambda_m = 3.6$ Å and $\lambda_h = 3.9$ Å, respectively, of methane in methanol. Lines are gaussian fits to the data.

6.4 Top Panel: QC contributions to $\mu^{ex}_w$ (Eq. 6.13) for methane in water as a function of the inner-shell radius, $\lambda_w$, for the water oxygens. Bottom Panel: QC contributions to $\mu^{ex}_m$ (Eq. 6.13) for methane in methanol as a function of the inner-shell radius, $\lambda_m$, for the methanol methyl groups. The corresponding radius for the hydroxyl groups of methanol is $\lambda_h = \lambda_m + 0.3$ Å. Symbols: $\mu^{ex}_w = \text{solid triangles}$; $kT \ln x_0 = \text{solid circles}$; $-kT \ln p_0 = \text{open circles}$; $\langle \epsilon \rangle = \text{solid squares}$; $\beta \sigma^2/2 = \text{open squares}$. Lines are drawn connecting the data points to guide the eye.
6.5 Methanol concentration dependence of the QC contributions to $\mu_{ex}$ for methane in aqueous methanol solutions evaluated using inner-shell radii of $\lambda_w = 3.3$ Å, $\lambda_m = 3.6$ Å, and $\lambda_h = 3.9$ Å. Symbols: $\mu_{ex}$ = solid triangles; $kT \ln x_0$ = solid circles; $-kT \ln p_0$ = open circles; $\langle \epsilon \rangle$ = solid squares; $\beta \sigma^2 / 2$ = open squares. Lines are linear fits to the data. Conditional binding energy probability distribution as a function of methanol concentration (Fig. A.3) and error estimates associated with the gaussian approximations for those probability distributions (Table B.1) are provided in the Appendix B.

6.6 Methane-methanol preferential interaction parameter $\Gamma_{sc}$, and the QC contributions to $\Gamma_{sc}$ (Eq. 6.13) for 10 mol% methanol/water solution as the function of inner-shell radius, $\lambda_w$, for water. The corresponding radii for the methyl and hydroxyl groups of methanol are defined as $\lambda_m = \lambda_w + 0.3$ Å and $\lambda_h = \lambda_w + 0.6$ Å, respectively. Symbols: $\Gamma_{sc}$ = diamonds; QC contributions from $kT \ln x_0$ = solid circles, $-kT \ln p_0$ = open circles, $\langle \epsilon \rangle$ = solid squares, and $\beta \sigma^2 / 2$ = open squares. Lines are drawn connecting the data points to guide the eye.

6.7 Preferential interaction parameter $\Gamma_{sc}$ for methanol with methane in aqueous methanol solutions at different methanol concentrations derived from KB theory as a function of the cutoff radius, $R$ (Eq. 6.4). Arrows on the right denote $\Gamma_{sc}$ at different methanol concentrations derived from the QC/PDT analysis (Table 6.3). The shaded region shows the range of inner-shell radii considered in that analysis. Methanol mole fractions are: 0.05 (solid line), 0.10 (dotted line), 0.20 (dashed line), 0.30 (dash-dot line).

7.1 Specific sites of our interest on lysozyme surface has been colored according to the atom type: cyan - carbon; red - oxygen, blue - nitrogen, white - hydrogen. The other heavy atoms of lysozyme are shown in orange.

7.2 Co-solvent preferential interaction parameter of lysozyme in aqueous solution of 2 mol% urea (solid line) and 10 mol% methanol (dashed line), evaluated by counting the number of co-solvent (center of mass) and water (oxygen) molecules in the vicinity of the entire protein surface (Eq. 7.3).
### 7.3 Local co-solvent preferential interaction parameters of the protein sites

in aqueous solution of 2 mol% urea (top panel) and 10 mol% methanol (bottom panel) evaluated by KB theory using smeared proximal radial
distribution functions of water (oxygen) and co-solvent (local smearing around each heavy atoms) molecules around the protein sites. Legends:
solid line = site 1; dashed line = site 2; dotted line = site 3; long dashed line = site 4; dash dotted line = site 5; long dash dotted line = site 6; dash dot dotted line = site 7.

### 7.4 Local co-solvent preferential interaction parameters of the protein sites

in aqueous solution of 2 mol% urea (top panel) and 10 mol% methanol (bottom panel) evaluated using QC theory (Eq. 7.7) as a function of the conditioning radius, $\lambda_w$, for water oxygens defining the inner shell. Legends: solid circle = site 1; open circle = site 2; solid square = site 3; open square = site 4; solid triangle up = site 5; open triangle up = site 6; open triangle down = site 7. Lines are drawn connecting the data points to guide the eye.

### 7.5 Comparison of the local co-solvent preferential interaction parameters of the protein sites in aqueous solution of 2 mol% urea (top panel) and 10 mol% methanol (bottom panel) evaluated using QC theory (Eq. 7.7), for a inner shell defined by water conditioning radius of $\lambda_w = 2.8$ Å, against its inner shell contribution and the values evaluated by KB theory using smeared radial distribution functions of water and co-solvent around the sites.

### 8.1 Zimm plots (Eq. 8.3) of light scattering from aqueous solutions containing lysozyme (1) and hPEG (2) as a function of lysozyme concentration at 6.2 pH and 0.8 M ionic strength (including 0.2 M sodium phosphate buffer and 0.5 M NaCl) relative to light scattering from hPEG solutions at the same solution conditions. Symbols correspond to hPEG concentrations (bottom to top) of 0, 50, 70, 90, and 110 mg/ml. Lines are linear fits to the data. The data at each hPEG concentration, starting with 50 mg/ml, are shifted upwards by 2 units for visual clarity.
8.2 Zimm plots (Eq. 8.3) of light scattering from aqueous solutions containing lysozyme (1) and mPEG (2) as a function of lysozyme concentration at 6.2 pH and 0.8 M ionic strength (including 0.2 M sodium phosphate buffer and 0.5 M NaCl) relative to light scattering from mPEG solutions at the same solution conditions. Symbols correspond to mPEG concentrations (bottom to top) of 0, 50, 70, and 90 mg/ml. Lines are linear fits to the data. The data at each mPEG concentration, starting with 50 mg/ml, are shifted upwards by 2 units for visual clarity.

8.3 Zimm plots (Eq. 8.3) of light scattering from aqueous solutions containing lysozyme (1) and hPEG (2) as a function of lysozyme concentration, at 7.0 pH and 0.01 M ionic strength (including 10 mM Bis-Tris buffer) adjusted using NaCl, relative to light scattering from hPEG solutions at the same solution conditions. Symbols correspond to hPEG concentrations (bottom to top) of 0, 30, 50, 70, 90, and 110 mg/ml. Lines are linear fits to the data. The data at each hPEG concentration, starting with 30 mg/ml, are shifted upwards by 2 units for visual clarity.

8.4 Zimm plots (Eq. 8.3) of light scattering from aqueous solutions containing lysozyme (1) and mPEG (2) as a function of lysozyme concentration, at 7.0 pH and 0.01 M ionic strength (including 10 mM Bis-Tris buffer) adjusted using NaCl, relative to light scattering from mPEG solutions at the same solution conditions. Symbols correspond to mPEG concentrations (bottom to top) of 0, 30, 50, 70, 90, and 110 mg/ml. Lines are linear fits to the data. The data at each mPEG concentration, starting with 30 mg/ml, are shifted upwards by 2 units for visual clarity.

8.5 The dependence of the apparent molecular weight of the protein (Eq. 8.5) on polymer concentration, in hPEG (2) or mPEG (2)-lysozyme (1) aqueous solutions at ionic strengths of 0.8 M (top figures) and 0.01 M (bottom figures). Left panels: hPEG solutions; Right panels: mPEG solutions.

8.6 The dependence of the apparent osmotic virial coefficient (Eq. 8.4) of the protein on polymer concentration, in hPEG (2) or mPEG (2)-lysozyme (1) aqueous solutions at ionic strengths of 0.8 M (top figures) and 0.01 M (bottom figures). Left panels: hPEG solutions; Right panels: mPEG solutions.
9.1 Zimm plots (Eq. 9.8) for aqueous solutions of lysozyme (2) and α-chymotrypsinogen A (3). Symbols correspond to the experimental data and the lines are linear fits, at each constant concentration of α-chymotrypsinogen A. Symbols represent different concentrations of α-chymotrypsinogen A: solid circles - 0 mg/ml; open circles - 0.45 mg/ml; solid squares - 0.89 mg/ml; open squares - 1.64 mg/ml; solid triangles - 2.57 mg/ml; open triangles - 3.49 mg/ml.

9.2 Zimm plots (Eq. 9.8) for aqueous solutions of lysozyme (2) and α-chymotrypsinogen A (3). Symbols correspond to the experimental data and the lines are the linear fits, at each constant concentration of lysozyme. Symbols represent different concentrations of lysozyme: solid circles - 0 mg/ml; open circles - 0.52 mg/ml; solid squares - 0.88 mg/ml; open squares - 1.68 mg/ml; solid triangles - 2.83 mg/ml; open triangles - 3.58 mg/ml.

9.3 The inverse apparent molecular weights of lysozyme (2) and α-chymotrypsinogen A (3) plotted as a function of the concentration of the second protein (Eq. 9.9) in the solution.

A.1 KB integrals of the solvent around various solutes in neat water (top panel) and in neat methanol (bottom panel), as a function of cutoff distance, $R_c$, from the solutes, determined using the smeared effective grand canonical distribution of solvent around the solutes: methane - solid line; ethane - dashed line; neopentane - dash and dotted line; TMA - dotted line. The radii of 2.3 Å and 3.3 Å are used for smearing the distributions of water and methanol, respectively.

A.2 Effect of concentration of methanol on the preferential interaction of methanol with neopentane (solid circles) and TMA (open circles). The lines are the linear fit to the data with $R^2 = 0.98$ for neopentane and $R^2 = 0.94$ for TMA.

A.3 Preferential interaction parameter of methanol with various solutes as a function of the cutoff distance, $R_c$, from the solutes in 50 mol % methanol solution calculated using the smeared effective grand canonical distributions of methanol and water. The different solutes are: methane - solid line; ethane - dotted line; neopentane - dashed line; TMA - dash-dotted line.
B.1 Conditional probability distributions of binding energies of methane in water when no waters are within: 3.3 Å (solid circles), 3.4 Å (open circles), 3.5 Å (solid squares), 3.6 Å (open squares), 3.7 Å (solid triangles), 3.8 Å (open triangles) of the center of the methane molecule. The energy distributions are successively shifted upwards by 2 units for visual clarity. Lines are gaussian fits to the data.

B.2 Conditional probability distributions of binding energies of methane in methanol when no methanol methyl group is within a distance of $\lambda_m$ and no methanol hydroxyl group is within $\lambda_h = \lambda_m + 0.3$ Å of the center of the methane molecule for $\lambda_m$: 3.6 Å (solid circles), 3.7 Å (open circles), 3.8 Å (solid squares), 3.9 Å (open squares), 4.0 Å (solid triangles), 4.1 Å (open triangles). The energy distributions are successively shifted upwards by 2 units for visual clarity. Lines are gaussian fit to the data.

B.3 Conditional probability distributions of binding energies of methane in aqueous methanol solutions at different methanol concentration when no water oxygens are within 3.3 Å, no methanol methyl groups are within 3.6 Å and no methanol hydroxyl groups are within 3.9 Å, of the center of the methane molecule. Methanol mole fractions are: 0.00 (solid circles); 0.05 (open circles); 0.10 (solid squares); 0.20 (open squares); 0.30 (solid triangles). Lines are gaussian fit to the data. The energy distributions are successively shifted upwards by 2 units for visual clarity.

C.1 Logarithm of conditional probability distribution of interaction energies of 7 protein sites (Table 7.1) in neat water (top panel) and in aqueous solution of 2 mol% urea (bottom left panel) and 10 mol% methanol (bottom right panel) such that no water or co-solvent molecule is within inner shell volume defined by water conditioning radius, $\lambda_w$, of 2.8 Å. The corresponding conditioning radii for heavy atoms of urea and methanol are given in Table 7.3. Symbols: protein site 1 = black circles; protein site 2 = red circles; protein site 3 = green circles; protein site 4 = blue circles; protein site 5 = violet circles; protein site 6 = dark green circles; protein site 7 = orange circles. Lines are the gaussian fit to the simulation data. Successive energy distributions are shifted upwards by 2 units for visual clarity. The average interaction energy of the protein site 7 (carbonyl group of Leu129) is shifted towards right by 22 kcal/mol to fit all distribution within the same graph.
C.2 Quasi chemical contributions to excess chemical potential of 7 protein sites in neat water as a function of water conditioning radius, $\lambda_w$. Symbols: $kT \ln x_0 = $ solid circles; $-kT \ln p_0 = $ open circles; $\langle \epsilon \rangle = $ solid squares; $\beta \sigma^2/2 = $ open squares; $\mu_{ex}^\alpha = $ open triangles. Lines are drawn connecting the data points to guide the eye. . . . . . . . . . . . . . . 181

C.3 Quasi chemical contributions to excess chemical potential of 7 protein sites in 2 mol% aqueous urea solution as a function of water conditioning radius, $\lambda_w$. The corresponding conditioning radii for the heavy atoms of urea are calculated as $\lambda_i = \lambda_w - 1.3 + R_i$. Symbols: $kT \ln x_0 = $ solid circles; $-kT \ln p_0 = $ open circles; $\langle \epsilon \rangle = $ solid squares; $\beta \sigma^2/2 = $ open squares; $\mu_{ex}^\alpha = $ open triangles. Lines are drawn connecting the data points to guide the eye. . . . . . . . . . . . . . . 182

C.4 Quasi chemical contributions to excess chemical potential of 7 protein sites in 10 mol% aqueous methanol solution as a function of water conditioning radius, $\lambda_w$. The corresponding conditioning radii for the heavy atoms of methanol are calculated as $\lambda_i = \lambda_w - 1.3 + R_i$. Symbols: $kT \ln x_0 = $ solid circles; $-kT \ln p_0 = $ open circles; $\langle \epsilon \rangle = $ solid squares; $\beta \sigma^2/2 = $ open squares; $\mu_{ex}^\alpha = $ open triangles. Lines are drawn connecting the data points to guide the eye. . . . . . . . . . . . . . . 183

C.5 Quasi chemical contributions to preferential interaction parameters of 7 protein sites in 2 mol% aqueous urea solution as a function of water conditioning radius, $\lambda_w$. The corresponding conditioning radii for the heavy atoms of urea are calculated as $\lambda_i = \lambda_w - 1.3 + R_i$. Symbols: $kT \ln x_0 = $ solid circles; $-kT \ln p_0 = $ open circles; $\langle \epsilon \rangle = $ solid squares; $\beta \sigma^2/2 = $ open squares; $\Gamma_{ac} = $ open triangles. Lines are drawn connecting the data points to guide the eye. . . . . . . . . . . . . . . 184

C.6 Quasi chemical contributions to preferential interaction parameters of 7 protein sites in 10 mol% aqueous methanol solution as a function of water conditioning radius, $\lambda_w$. The corresponding conditioning radii for the heavy atoms of methanol are calculated as $\lambda_i = \lambda_w - 1.3 + R_i$. Symbols: $kT \ln x_0 = $ solid circles; $-kT \ln p_0 = $ open circles; $\langle \epsilon \rangle = $ solid squares; $\beta \sigma^2/2 = $ open squares; $\Gamma_{ac} = $ open triangles. Lines are drawn connecting the data points to guide the eye. . . . . . . . . . . . . . . 185

xxiv
CHAPTER 1

Introduction

Protein-protein interactions are the hallmark of virtually all biological processes. In portraying the diverse and vital roles of proteins in the cellular processes, proteins have been characterized as "the working horses of the cellular machinery" \[2\]. The affinity and specificity of protein-protein interactions provide the basis for classifying these interactions as strong interactions if the disassociation constant, $K_D < 10^{-6}M$ (e.g., obligatory or permanent protein complexes) or weak interactions for $K_D > 10^{-6}M$ (e.g., most transient protein complexes) \[3\]. In addition to specific functional interactions, weak and non-specific interactions occur between various proteins in the crowded cellular environment \[4, 5\]. Under unfavorable physiological conditions, interactions leading to protein aggregation and amyloid plaque formation can also occur \[6\]. The formation of amyloid fibrils have been identified as the key pathological condition for all neuro-degenerative diseases, such as Alzheimers disease, Parkinsons disease, Huntingtons disease \[7\]. Knowledge of protein-protein interactions is, therefore, foundational to the understanding of cellular mechanisms.

Protein-protein interactions are also the key elements in molecular thermodynamic models of biochemical processes, most notably the commercial production of therapeutics. Protein therapeutics are gaining popularity in effectively treating diseases
such as diabetes, anemia, hepatitis and cancer, due to their high specificity and lower probability of eliciting immune responses. To date, more than 130 different proteins or peptides have been approved for clinical use by the US Food and Drug Administration (FDA) [8]. For bulk production of high purity protein drugs to meet FDA standards, pharmaceutical industries typically process proteins by repeated precipitations and crystallizations. The crystalline form of a protein drug is easy to store, has improved shelf life, and can facilitate controlled drug delivery. Protocols for protein crystallization are routinely determined by trial and error, as numerous factors like ionic strength of the protein solution, concentration of precipitants, pH, and temperature can play important roles in protein crystallization [9].

The identification of solution conditions or the so-called “crystallization slot” [10] in which proteins readily crystallize promises a more systematic approach. This crystallization slot is defined by the protein osmotic second virial coefficient. When this osmotic second virial coefficient falls within the range of \(-1 \times 10^{-4}\) to \(-8 \times 10^{-4}\) mol ml/g\(^2\) crystallization occurs, and when this osmotic second virial coefficient is less than \(-8 \times 10^{-4}\) mol ml/g\(^2\), precipitation occurs [10]. At the molecular level, this osmotic second virial coefficient characterizes protein-protein interactions. Therefore, knowledge of these interactions is the foundation for protein solution thermodynamic models capable of predicting conditions for the successful crystallization of proteins in commercial processes.

Protein crystallization is typically induced by adding a precipitating agent or cosolvent, such as inorganic salts or organic compounds: alcohols, polyols, or polyethylene glycol (PEG) [9 11 12]. Recipes for crystallizing over 2100 crystal forms of 1400 macromolecules are accessible in the NIST/NASA/CARB Biological Macromolecule
Crystallization Database, version 3.0 [12, 13]. Over the past two decades, a number of research groups have developed models of protein-protein interactions in aqueous salt solutions by treating protein solutions as colloidal suspensions. In many aspects, protein solutions resemble colloidal suspensions: the solutes are charged Brownian objects with sizes between a few nanometers and a few tens of nanometers. Hence, the classical Derjaguin-Landau-Verwey-Overbeek (DLVO) theory [14] developed to describe interactions between charged colloidal surfaces in a liquid medium has been applied to describe protein-protein interactions in solution [15, 16, 17, 18]. Primitive models [16, 17, 19] based on DLVO theory, however, neglect the spatial heterogeneity of protein surfaces, and therefore, cannot predict the orientational dependence of protein-protein interactions.

A more realistic model of protein-protein interactions must consider the orientational dependence of these interactions to accurately predict osmotic second virial coefficients [20, 21]. Specifically, these interactions are dominated by protein-protein configurations that have high complementarity [20, 21]. Specific hydration of the protein surface is uneven due to spatial heterogeneities in protein chemistry and molecular topography of the surface. Uneven, preferential hydration across the protein surface eliminates some complementary configurations, which alters the protein-protein interactions [22]. In dilute electrolyte solution, models for protein-protein interactions incorporating the strongly associated waters as an integral part of protein are in quantitative agreement with experimentally measured osmotic second virial coefficients [23, 22]. However, the molecular mechanism of protein precipitation by organic co-solvents, such as alcohols and polyols or low molecular weight PEG, is still unresolved. The free energy of transferring a protein from neat water to aqueous
solution of co-solvent is correlated to the preferential interaction of co-solvent with the protein \[24, 25, 26\]. A detailed understanding of protein hydration and co-solvent preferential interactions are, therefore, indispensable for modeling protein-protein interactions, and ultimately the osmotic second virial coefficients of proteins in solution, from which solution free energies can be derived.

The principal objective of the work presented in the thesis is to understand the fundamental, molecular-level interactions among proteins, water and co-solvents – specifically, alcohols and polyols – in exploring the role of hydration and co-solvent preferential interactions on protein-protein interactions.

The quasi-chemical theory \[27, 28, 29\] provides an appropriate molecular framework for examining protein hydration and preferential interactions at specific protein sites. Motivated by a quasi-chemical (QC) view of hydration, Asthagiri, Paliwal, et al., \[23, 22\] define specific hydration sites on the surface of globular proteins in terms of the local water density at each site relative to bulk water density. Characterization of water association on the basis of water dynamics near the protein surface is an alternative approach \[30\]. From the dynamics of water molecules at hydration sites with high occupancies, chapter 2 identifies two kinetic regimes: long residence times relative to vacancy times for a single water molecule, and short residence times with high turnover involving multiple water molecules. Water molecules with long residence times tend to be buried in regions of high surface roughness and charge heterogeneity; thus, they have less impact on protein-protein interactions compared to those hydration sites, which are more solvent accessible.

Chapter 3 presents the hydration characteristics of concave and convex structure of a large apolar sinusoidal surface. The effect of water-surface atom interactions is more
prominent in concave regions than on convex regions; indeed, the crevices on protein surfaces are mostly concave pockets. Chapter 4 discusses the hydration of amino acid side chains and the applicability of group contribution on protein hydration and its partial molar volume. The local water density around the amino acids are obtained from the proximal distributions of water about its constituent molecular groups.

The molecular preferential interaction of co-solvent with the solute, determining the solubility of the solute in the solution, can be described based on Kirkwood-Buff (KB) theory [31, 32]. However, long range oscillations in the radial distribution functions lead to poor convergence of the KB integrals when these distribution functions are derived from molecular simulations. Two alternative approaches to determine co-solvent-solute preferential interaction are discussed in chapters 5 and 6. In the first approach (chapter 5), a modified KB theory [33] is developed in which oscillations in KB integrals, which lead to poor convergence, are damped by smearing [34, 35] the corresponding radial distribution functions. Smearing assumes a distribution of co-solvent molecular reference points have equal probability, rather than a single reference point, such as co-solvent center of mass. In the second approach (chapter 6), preferential interactions are estimated through quasi chemical theory. In chapter 7, the quasi chemical theory is applied to protein solutions to evaluate local preferential interactions at specific sites on protein surface.

Low-molecular weight poly(ethylene glycol) (PEG) is the most commonly used additive for inducing protein phase separation [9, 13, 36, 37, 38]. Protein interactions with co-solvents like PEG or other globular protein in the solution, are embodied in cross osmotic virial coefficients, and are measured through static light scattering. Chapter 8 discusses the effects of PEG end-group hydrophilicity/hydrophobicity on
PEG-lysozyme and lysozyme-lysozyme interactions. The cross interactions between two globular proteins – lysozyme and α-chymotrypsinogen A – are reported in chapter 9. Finally, conclusions and recommendations for future work are presented in chapter 10.
2.1 Introduction

Configurational complementarity in protein-protein interactions is a hallmark of molecular recognition, and leads naturally to a consideration of the molecular nature of protein hydration \[20, 21\]. The essential features of complementarity embodied in protein structures can be captured to some extent using continuum solvent models. However, water molecules strongly associated with the protein make important contributions by occupying space or providing hydrogen bond donors and acceptors at the protein-water interface that constrain conformational space, and these effects are lost if the bound water molecules are not explicitly taken into account.

In the quasi-chemical (QC) view of protein hydration, it is natural to consider strongly associated water molecules as part of the protein \[27, 28, 29\]. Protein solution thermodynamics is then modeled in terms of quasi-components comprised of the protein and associated water molecules immersed in a statistical field due to the remaining solvent medium. We implemented this QC view in previous work by defining specific hydration sites near the protein surface characterized by high water occupancies \[22\]. An important finding of our earlier study was to show that the
spatial distribution of these hydration sites plays an essential role in determining the configurational complementarity of protein-protein interactions.

Specific hydration sites in that study were characterized by their local water densities, which we expressed in terms of the logarithm of the chemical equilibrium constant, \( \eta \), for water partitioning at a specific site relative to bulk water,

\[
\eta \equiv \ln \left( \frac{\rho}{\rho_b} \right) = -\beta \Delta \mu_w^{ex}.
\]  

(2.1)

Here \( \Delta \mu_w^{ex} \) is the excess chemical potential of water at the hydration site relative to bulk water, \( \rho/\rho_b \) is the corresponding ratio of water densities, and \( \beta^{-1} = kT \), the thermal energy. Eq. 2.1 provides a thermodynamic framework for selecting hydration sites representing bound water molecules based on explicit-water molecular dynamics (MD) simulations that supply the required densities. Our criterion for strongly associated or bound water was \( \eta > 2 \), which corresponds to a local density more than seven times that of bulk water, or roughly that found for the maximum density in the first hydration shell of simple monovalent or divalent ions [39, 40]. Relaxing this criterion to include a larger number of more weakly associated water molecules was found to have a minimal effect on the osmotic second virial coefficient for protein-protein interactions in dilute aqueous solution [22]. Thus, \( \eta > 2 \) defined a lower bound on the number of explicit water molecules that must be considered in order to obtain the full effect of water association on these protein-protein interactions.

Water association can also be characterized by water dynamics near the protein surface. Experimental techniques such as magnetic resonance dispersion [41], nuclear magnetic resonance [42] and quasi elastic neutron scattering [43] measure relaxation times for waters buried in the interior of proteins and residence times for waters within
the first hydration shell. These times are typically on the order of 10 ns - 1 ms and 10 ps - 1 ns, respectively for small globular proteins. Much faster water relaxation times (0.5 ps - 100 ps) in the vicinity of a surface tryptophan residue are probed by fluorescence spectroscopy [44]. Based on these timescales, water molecules have been broadly categorized as: (i) internal water ($\tau \sim 10$ ns to ms), (ii) water molecules that interact with the protein surface ($\tau \sim 10$ ps - 100 ps), and (iii) bulk water ($\tau \sim 1$ ps) [41].

Water dynamics near protein surfaces has also been extensively investigated by MD simulations, and characterized by mean residence times derived from various correlation functions. The survival probability correlation function [45, 46, 47, 48, 49] is one such correlation function that has been widely used. The survival probability is defined as the probability of finding a water molecule within a region of interest; e.g., the first hydration shell, for a specific period of time. The hydrogen bond correlation function [50] and the solvation energy correlation function [51] have also been used. The hydrogen bond correlation function is defined in terms of the probability of finding a water molecule hydrogen bonded to a protein atom for a specific period of time, with the hydrogen bond typically defined by the donor-acceptor distance and angle. The solvation correlation function is defined in terms of fluctuations in the potential energy of solute-solvent interactions.

In each case, the correlation function derived from a MD simulation is fit to either a single exponential or a series of exponential functions, or a stretched exponential function [45, 46, 47, 48, 49, 50, 51]. A single exponential fit gives the water residence time directly. For a series of exponential functions, the mean water residence time is calculated as the weighted average of the characteristic time constants. While for a
stretched exponential function, the mean residence time is defined in terms of both the characteristic time and the stretched exponent.

It is widely accepted that molecular features of the protein surface influence water dynamics. However, previous attempts to derive a correlation between water dynamics and local chemical heterogeneities of the protein surface have produced results that are inconclusive or even contradictory [45, 48, 52, 49, 47, 50]. For example, the MD simulation study of water dynamics near copper plastocyanin [45] and crambin [47] found that the mean residence times computed from the survival probability correlation function depend on the chemical nature of proximal amino acids with 

\[ \tau_{\text{charged}} \geq \tau_{\text{polar}} > \tau_{\text{non polar}} \approx \tau_{\text{bulk}}. \]

A similar ordering of water residence times was observed for all 20 amino acids in the end capped AXA tripeptide motif [52]. However, an entirely different dependence of the water survival time was reported in MD simulation study of bovine pancreatic trypsin inhibitor [49]:

\[ \tau_{\text{polar}} > \tau_{\text{non polar}} > \tau_{\text{charged}}. \]

MD simulations have also shown that both the survival probability time for waters around negatively charged residues [48, 47] and the time that water molecules remain hydrogen bonded with negatively charged residues [45] are significantly longer than that for waters near positively charged residues. In contrast, the hydrogen bond correlation time for water near positively charged residues was found to be higher than that near negatively charged residues in the MD simulation study of HP-36 [50].

The correlation between local water densities and the chemical nature of proximal amino acids appears to be even weaker than that between water residence times and the local chemical environment [52], suggesting at best a weak correlation between local water densities and residence times. Indeed, the lack of a correlation was observed between water densities and water residence times around the constituent amino acids.
in a MD simulation study of myoglobin [46]. In this study, water was found to reside longer in clefts and notches of the protein surface irrespective of the local chemical environment, suggesting that local topographical features of the protein surface may be the dominant factor influencing water dynamics on the protein surface.

Here we analyze protein hydration described by the dynamics of water association with the protein surface, and compare this description to a thermodynamic description based on the QC view of protein hydration. Our interest in the spatial distribution of protein hydration naturally leads us to consider the relationship between the spatial heterogeneity of water dynamics near the protein surface and the local topography and the chemical composition of the protein surface.

2.2 Theory

Our analysis of water dynamics is based on specific hydration sites near the protein surface, and uses a master equation to describe the chemical reaction dynamics for transitions between occupied and unoccupied states of these sites [53, 54]. The rate constants in this two-state model are related to quantities that can be extracted directly from the MD simulations: the average time, $\tau_1$, a site is occupied by a water molecule, and the average time, $\tau_0$, that site remains unoccupied. The resulting probability of finding $n$ occupied states out of a total number of $N$ realizations is given by the binomial distribution,

$$P_n = \frac{N!}{n!(N-n)!} q^n (1-q)^{N-n} \tag{2.2}$$

with $q = \tau_1 / (\tau_1 + \tau_0)$. The mean of this distribution, $\langle n \rangle$, gives the local water density at a site of unit volume,

$$\rho \equiv \frac{\langle n \rangle}{N} = \frac{\tau_1}{\tau_1 + \tau_0}. \tag{2.3}$$
Defining the average cycle time, $\tau_{cyc} = \tau_1 + \tau_0$, and substituting for $\eta$ in Eq. 2.1 provides the desired relationship between the local water density and the average occupancy and vacancy times at each hydration site:

$$\eta = \ln \left( \frac{\tau_1}{\tau_{cyc}} \right) - \ln \left( \frac{\tau_1}{\tau_{cyc}} \right)_{bulk} . \tag{2.4}$$

The ratio for bulk water is fixed by the density of water at the conditions of interest (Eq. 2.3). For a hydration site 1 Å³ in volume and water at 300 K, we have

$$\eta = 3.4 + \ln \left( \frac{\tau_1}{\tau_{cyc}} \right) . \tag{2.5}$$

The maximum value of $\eta = 3.4$ is obtained when

$$\tau_1 \to \tau_{cyc} = \tau_1 + \tau_0 ,$$

and can be realized by two different kinetic pathways. A site can be occupied by a single water molecule for long periods of time relative to the time it is unoccupied – *i.e.*, $\tau_1 \gg \tau_0$ – or it can be occupied for relatively short periods of time with high turnover – *i.e.*, $\tau_0 \to 0$ at small $\tau_1$. These two pathways distinguish sites with bound waters from those sites that are simply highly accessible to water. Of course, a practical definition of kinetically bound water sites will depend on the extent to which $\tau_1$ is taken to be greater than $\tau_0$. However, the least restrictive kinetic criterion for water association in either case is $\tau_1 > \tau_0$, which corresponds to $\eta > 2.7$.

We also note that the second moment of the binomial distribution, Eq. 2.2, relates local water density fluctuations to the fractional vacancy time at each hydration site,

$$\frac{\langle n^2 \rangle - \langle n \rangle^2}{\langle n \rangle} = \frac{\tau_0}{\tau_1 + \tau_0} . \tag{2.6}$$
Thus, the average occupancy and vacancy times of specific hydration sites in this kinetic description of preferential hydration includes information on both water densities and fluctuations in water densities locally near the protein surface.

2.3 Methods

Lysozyme (PDB ID: 1LYZ) [55] was solvated in a cubic box 62 Å on a side containing 7107 TIP3P water molecules [56] and staphylococcal nuclease (PDB ID: 1JOO) [57] was solvated in a cubic box of 82 Å containing 18480 TIP3P water molecules. MD simulations of these proteins were carried out at 300 K and 1 bar using NAMD 2.6 [58] with the CHARMM27 force field [59]. The average water residence time in the first hydration shell of proteins is on the order of a few picoseconds. We can consider a protein to be essentially rigid on this time scale; therefore, the protein atoms were held fixed throughout the simulation. Bulk water properties were determined from an independent MD simulation of 512 TIP3P water molecules at the same temperature and pressure.

Temperature was held constant in these simulations by applying Langevin dynamics to all heavy atoms using a damping coefficient of 1 ps$^{-1}$. Constant pressure was maintained using a Nosé-Hoover Langevin piston with a period of 200 fs and a decay of 100 fs. Periodic boundary conditions were imposed, and the particle-mesh Ewald method with a real-space cutoff of 12 Å was used in computing the electrostatic interactions. The same cutoff was applied to nonbonded nonelectrostatic interactions. The TIP3P water geometry was constrained by the SHAKE algorithm [60]. The system
was initially minimized for 20,000 steps and then equilibrated for 200 ps. Configurations were saved every 0.1 ps over a production run of 2 ns with a time step of 2 fs.

Specific hydration sites were defined as before [22] by constructing a network of grid points separated by 1 Å to fill the proximal volume within 3.5 Å of the heavy atoms on the protein surface. Water occupancy and vacancy times for each site were recorded over the course of the MD simulation, and arithmetic averages computed for both characteristic times. Different studies adopt different approaches to compute the average water occupancy or residence times. In most cases, the average residence time is estimated by fitting a series of exponential functions or a stretched exponential function to a time correlations function. The average residence times obtained from such fits are biased by the few infrequent long times a water molecule resides in the region of interest. By adopting the arithmetic averaging here, we obtain a more realistic representation of the frequency of water exchanges at a hydration site.

Using these methods, the number of hydration sites with $\eta > 2$ obtained from the MD simulation of lysozyme was 150 out of 8290 sites. This number is slightly higher than that reported previously – 135 out of 7855 sites [22] – and is attributed to the different reference frames that were used in constructing the network of grid points around the protein in the two studies. The pattern of high-occupancy hydration sites ($\eta > 2$) around the protein was found by visual inspection to be the essentially same, however, independent of the reference frame. For staphylococcal nuclease, 224 out of 12,936 hydration sites were found with $\eta > 2$. 
Figure 2.1: Natural logarithm of the ratio of average occupancy times ($\nu_1 = \tau_1/\tau_{1,b}$), average cycle times ($\nu_2 = \tau_{cyc,b}/\tau_{cyc}$), and the combination of the two times ($\nu = \nu_1\nu_2$), defined in Eq. 2.4 as a function of $\eta$ (Eq. 2.1) for hydration sites within 3.5 Å of the surface of lysozyme. Only those sites corresponding to $\eta > 2.0$ are shown. These characteristic times were obtained from MD simulations of lysozyme in TIP3P water at 300 K and 1 bar. Correlation coefficients for the linear fits of the data are 0.654, 0.176 and 0.999, respectively.
2.4 Results and Discussion

Fig. 2.1 confirms the relationship given by Eq. 2.4 between the water density at a specific hydration site and the average site occupancy and cycle times. The results show that the local water density is only weakly correlated with the average occupancy time, although a slightly better correlation is obtained for the occupancy time compared to the cycle time. This observation is in agreement with the findings of the MD simulation study of myoglobin in which no correlation was found between local water densities and residence times at specific hydration sites around this protein [46].

Plots of the average occupancy time, $\tau_1$, versus the average vacancy time, $\tau_0$, for all hydration sites on the surface of lysozyme and for the high occupancy sites ($\eta > 2.0$) on the surface of staphylococcal nuclease are shown in Figs. 2.2 and 2.3, respectively. Only 23 of the 150 high occupancy sites on lysozyme and 25 of the 224 high occupancy sites on staphylococcal nuclease are found below the $\tau_1 = \tau_0$ diagonal line, and as such, satisfy the kinetic criterion for bound water. These sites are also isolated from one another for the most part, as determined by computing clusters of sites that are within 4 Å of one another. For lysozyme, we obtained 18 clusters for the 23 sites with only a single cluster containing a maximum of three sites, while for staphylococcal nuclease, we obtained 23 clusters for the 25 sites with no cluster having more than two sites. The weak correlation between site water densities and residence times follows directly from the observation that a large fraction of the high occupancy sites are characterized by high turnover of water occupancies with occupancy times that span a relatively narrow range: $\tau_1 \sim 0.3 - 0.4$ ps, or roughly twice that for a site in bulk water (0.18 ps).
Figure 2.2: Average occupancy time, $\tau_1$, and average vacancy time, $\tau_0$, for each point in a grid filling the proximal volume within 3.5 Å of the heavy atoms on the surface of lysozyme, obtained from MD simulation of this protein in TIP3P water at 300 K and 1 bar. Solid triangles: $\eta > 2.7$; open circles: $2.7 > \eta > 2.0$; gray crosses: $\eta < 2.0$. The dashed $\tau_1 = \tau_0$ line separates sites that correspond to kinetically bound waters (solid triangles) from the sites corresponding to high water occupancies ($\eta > 2.0$) (open circles), while the dash-dot line separates the high occupancy sites from all other hydration sites (gray crosses). The inset shows only those data for $\tau_1, \tau_0 \leq 2$ ps.
Figure 2.3: Average occupancy time, $\tau_1$, and average vacancy time, $\tau_0$, for each point in a grid filling the proximal volume within 3.5 Å of the heavy atoms on the surface of staphylococcal nuclease, obtained from MD simulation of this protein in TIP3P water at 300 K and 1 bar. Solid triangles: $\eta > 2.7$; open circles: $2.7 > \eta > 2.0$; gray crosses: $\eta < 2.0$. The dashed $\tau_1 = \tau_0$ line separates sites that correspond to kinetically bound waters (solid triangles) from the sites corresponding to high water occupancies ($\eta > 2.0$) (open circles), while the dash-dot line separates the high occupancy sites from all other hydration sites (gray crosses).
Figure 2.4: Ten highest occupancy sites on the surfaces of spheres with the following chemical and/or topographical features: (a) smooth dipolar surface, (b) smooth random surface, (c) rough apolar surface, and (d) rough random surface. The four spheres have a radius of 10 Å, and are filled with neutral atoms in the interior. Atoms on the surface are: neutral (green), negatively charged (-0.51 e, red), or positively charged (+0.51 e, blue). The models are constructed to be electrostatically neutral by placing an atom of opposite charge below the surface for every charged atom on the surface. The gray spheres represent water oxygens at these high-occupancy sites. Lennard Jones parameters for all atoms are $\sigma = 3.4$ Å and $\epsilon = -0.12$ kcal/mol.
Figure 2.5: Average occupancy time, $\tau_1$, and average vacancy time, $\tau_0$, for the ten highest occupancy sites on the surface of four idealized (spherical) models of globular proteins obtained from MD simulations in TIP3P water at 300 K. Circles: smooth dipolar surface; diamonds: smooth random surface; crosses: rough apolar surface; triangles: rough random surface. The dashed line and the dash-dot line are defined in Figs. 2.2 and 2.3.
Different regimes of kinetic behavior on the $\tau_1-\tau_0$ plots of Figs. 2.2 and 2.3 can be related to local chemical and topographical features of protein surfaces by considering the four idealized models of protein surfaces depicted in Fig. 2.4. These models were chosen collectively to include an overall composition of surface charges (Figs. 2.4b and 2.4d) and surface roughness (Figs. 2.4c and 2.4d) representative of small globular proteins, in general. The corresponding $\tau_1-\tau_0$ plot for the ten highest occupancy sites (highest $\eta$ values) in each case is shown in Fig. 2.5. The overall range of $\tau_1-\tau_0$ values in this plot is strikingly similar to that obtained for lysozyme and staphylococcal nuclease (Fig. 2.2 inset and Fig. 2.3), although the range of $\tau_1-\tau_0$ values for the individual models are much more restricted. The similarity suggests that the hydration of these four idealized protein surfaces encompasses the different kinetic regimes of hydration behavior observed for lysozyme and staphylococcal nuclease.

For the smooth dipolar surface in Fig. 2.4a (no surface roughness), local heterogeneities exist only in the equatorial region separating the hemispheres of neutral and negatively charged atoms. It is striking that the highest occupancy sites are found only at this interface, rather than within the hemisphere of charged atoms. The same behavior is observed when the two hemispheres are neutral and positively charged or positively and negatively charged (not shown), indicating that the local water density is sensitive to local heterogeneities in surface charge, rather than the magnitude of the surface charge density. However, none of the high occupancy sites correspond to kinetically bound waters; i.e., $\tau_0 \gg \tau_1 \sim 0.2 - 0.3$ ps (Fig. 2.5). These sites also fall outside the range for high occupancy sites ($\eta > 2.0$) for lysozyme and staphylococcal nuclease due to the high circumferential mobility of water molecules in the equatorial region.

21
When the charges are dispersed over the surface (Fig. 2.4b), the average vacancy time is reduced significantly without much impact on the average occupancy times for the highest occupancy sites (Fig. 2.5). Nonetheless, $\tau_1 < \tau_0$ for all these sites, indicating that heterogeneities in the surface charge alone do not produce sites corresponding to kinetically bound waters. Virtually identical kinetic behavior is observed when local heterogeneities in surface roughness are introduced (Fig. 2.4c). However, the highest occupancy sites are now found in clefts and grooves on the rough surface.

It is only when the surface is rough and the surface charge is dispersed (Fig. 2.4d) that we find the highest occupancy sites corresponding to $\tau_1 > \tau_0$ (Fig. 2.5). In this case, $\tau_0 \sim 0.2 - 0.3$ ps and $\tau_1$ spans the same range of values obtained for lysozyme and staphylococcal nuclease (Fig. 2.2 inset and Fig. 2.3). We conclude, therefore, that local heterogeneities in both the surface charge and roughness are necessary to obtain specific hydration sites on the protein surface that correspond to kinetically bound waters.

We also calculated osmotic second virial coefficients for protein-protein interactions using different characterizations of preferential hydration defined by the kinetically bound water sites alone or by all the high occupancy sites. The results are shown in Table 2.1. The interaction part of the second virial coefficient, $\beta_{22}$, calculated here accounts for nonideal contributions due to protein-protein interactions, and is obtained by subtracting the Donnan contribution. Details of this calculation using a molecular thermodynamic model of protein solutions are given elsewhere [22, 23].

In this previous work, we showed that a number of highly complementary protein-protein contact configurations are eliminated by including a spatially heterogeneous distribution of explicit water molecules strongly associated with the protein surface.
Table 2.1: Interaction part of second virial coefficient, $\beta_{22}$, as a function of hydration conditions for lysozyme at pH 7 and ionic strength 0.007 moles/liter and staphylococcal nuclease at pH 6.5 and ionic strength 0.01 moles/liter

<table>
<thead>
<tr>
<th>Hydration condition</th>
<th>No. of sites</th>
<th>$\beta_{22} \times 10^4$ mol ml/g$^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lysozyme</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>continuum solvent</td>
<td>0</td>
<td>-87.44</td>
</tr>
<tr>
<td>kinetically bound waters sites ($\eta &gt; 2.7$)</td>
<td>23</td>
<td>-86.81</td>
</tr>
<tr>
<td>high occupancy sites ($\eta &gt; 2.0$)</td>
<td>150</td>
<td>-82.92</td>
</tr>
<tr>
<td><strong>Staphylococcal nuclease</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>continuum solvent</td>
<td>0</td>
<td>-80.58</td>
</tr>
<tr>
<td>kinetically bound waters sites ($\eta &gt; 2.7$)</td>
<td>25</td>
<td>-80.87</td>
</tr>
<tr>
<td>high occupancy sites ($\eta &gt; 2.0$)</td>
<td>242</td>
<td>-64.54</td>
</tr>
</tbody>
</table>

at specific hydration sites. Short-ranged protein-protein interactions thus become less favorable. This effect is seen in Table 2.1 by comparing $\beta_{22}$ calculated for a continuum solvent (no explicit water molecules) to that calculated with all high occupancy hydration sites taken into account. For both lysozyme and staphylococcal nuclease, $\beta_{22}$ becomes less negative – less favorable protein-protein interactions – when explicit waters at the high occupancy hydration sites are taken into account. In contrast, the contribution from the kinetically bound water sites alone is negligible; the calculated $\beta_{22}$ is essentially the same as that for the continuum solvent. The observation reflects our finding that explicit water molecules located at the kinetically bound hydration sites in regions of high surface roughness and charge heterogeneity tend to be buried; thus, they have less impact on reducing the complementarity of protein-protein contact configurations compared to explicit water molecules located at the other high occupancy hydration sites, which are more solvent accessible.
2.5 Conclusions

Preferential hydration of protein surfaces was analyzed by computing average water residence times and vacancy times at specific hydration sites on the protein surface. This analysis revealed two distinct kinetic regimes for those hydration sites defined thermodynamically to have high local water densities: long residence times relative to vacancy times for a single water molecule, corresponding to kinetically bound water molecules, and short residence times with high turnover involving multiple water molecules. Those sites corresponding to kinetically bound water molecules comprise only a small fraction of the total number of high occupancy sites, and are correlated with local heterogeneities in both surface charge and roughness. Moreover, these sites have little impact on calculated osmotic second virial coefficients for protein-protein interactions. The impact of preferential hydration on these weak protein-protein interactions is due primarily to the preferential hydration of sites characterized by high occupancy and high turnover — i.e., those sites on the protein surface that are accessible to water.

In deriving a relationship between kinetic and thermodynamic views of preferential hydration (Eq. 2.4), we found that the thermodynamic characterization in terms of the local water density at specific hydration sites and the kinetic characterization in terms of water occupancy and vacancy times at these sites are not equally informative. Specifically, while it is possible to obtain the local water density from a knowledge of site occupancy and vacancy times, it is not possible to derive the average occupancy and vacancy times knowing just the local densities for the hydration sites. Indeed, the weak correlation that we found between local water densities and average residence times follows directly from this analysis and the observation that most high occupancy
sites on the two protein surfaces we studied have a narrower range of occupancy times compared to the range of vacancy times.

The thermodynamic and kinetic perspectives of preferential hydration would be equally informative if just the ratio of water occupancy and vacancy times at each site was sufficient for an accurate description of hydration. We find, however, two distinct regimes of kinetic behavior for lysozyme and staphylococcal nuclease, and collectively, for the four spherical models of a protein surface – one characterized by specific hydration sites with high turnover in occupancies ($\tau_1 \sim \text{constant} < \tau_0$), and the other characterized by specific hydration sites with strongly associated or kinetically bound waters ($\tau_0 \sim \text{constant} < \tau_1$). We conclude, therefore, that a more complete description of the preferential hydration of protein surfaces is achieved when occupancy and vacancy times are taken to be independent of one another.

Of course, the corollary is that an additional parameter in the thermodynamic analysis is required to obtain an equivalent description of preferential hydration. Recognizing that the average site occupancy and vacancy times in the kinetic model characterize both water densities (Eq. 2.3) and fluctuations in water densities (Eq. 2.6) locally near protein surfaces, we submit that the logical, although not necessarily practical choice for an additional thermodynamic parameter is the water-water pair distance distribution function. Extracting this parameter from MD simulations in the heterogeneous environment of the protein-water interface with the spatial resolution demonstrated here for $\tau_0$ and $\tau_1$ is a formidable, if not impossible task. An advantage of the kinetic model is that water occupancy and vacancy times characteristic of specific hydration locally at sites on the protein surface are indeed readily accessible from MD simulations.
Finally, the kinetic analysis of preferential hydration presented here does not take into account any coupling of water dynamics to the protein dynamics, since the protein was held fixed in our MD simulations. This coupling undoubtedly would be important in an analysis of water dynamics near protein surfaces. In the context of our focus on protein hydration, though, we note that, the water occupancy times corresponding to the high occupancy/high accessibility sites are all less than 0.5 ps, which is more than an order of magnitude smaller than the characteristic time for side chain rotations of the amino acids on a protein surface [61]. We conclude, therefore, that our kinetic characterization of preferential hydration is unaffected by protein dynamics on these longer time scales, other than introducing the need to consider an ensemble of protein configurations that would be accessible on time scales for protein-protein interactions.
CHAPTER 3

A Simulation Study on the Wetting Characteristics of a Sinusoidal Hydrophobic Wall

3.1 Introduction

Hydrophobic interaction is one of the main driving forces for biological self-assembly processes, such as protein folding and interactions, membrane formation and molecular recognition [62, 63, 64, 65]. Our current understanding of the hydrophobic effects derives from the classic theories like scaled-particle theory [66, 67, 68], Pratt-Chandler integral-equation theory [69], configurational entropy expansion [70], density fluctuations [71], information theory based cavity occupancy fluctuations [72]. Except for the scaled-particle theory, other theories describe hydrophobic effects at molecular length scale. Biological interactions, however, occur over a spectrum of length scales ranging from amino acids to large proteins surfaces. The hydrophobic effect at various length scales are quite different [73, 74]. Extrapolation of the molecular level hydrophobic interactions to mesoscopic organization of proteins and biomembranes, therefore, may not be accurate.

The length scale dependent hydrophobic effect originates from the associative nature of water [75]. Small solutes can fit into the hydrogen bond network of water without much perturbations. With increase in solute size, water molecules are forced
to adopt more restrictive orientations to maximize the number of hydrogen bonds. From simulations it was computed that the orientational entropy per water molecule in the first hydration shell of the solutes exhibits a maximum at a effective solute diameter of 4 Å \[76\]. In fact, a water molecule over a hydrophobic surface effectively loses one H-bond resulting in dangling hydrogen bonds \[77\]. The cross-over of small and large scale hydrophobicity apparently occurs at subnanoscopic length of \(\sim 5 \text{ Å} \[78, 79\], comparable to the size of a neopentane molecule.

The loss in water-water interactions as a water molecule approaches a hydrophobic surface results in an effective cavity expulsion potential \[80\]. Using a perturbation model for solute-water potential of mean force, Hummer and Garde computed the effective cavity expulsion potential for LJ solutes upto 7.5 Å in diameter from molecular simulations. For small solutes, diameter less than 4 Å, the cavity expulsion potential is negligible. In contrast, for larger solute the cavity expulsion potential increases progressively with increase in solute size, resulting in a weak dewetting of the solute. The cavity expulsion potential depends on the curvature of the hydrophobic, for spherical solutes (convex surface – positive curvature) it is centered inside the solute, where as for the concave pockets (negative curvature) it can be outside the nonpolar surface. They found the cavity expulsion potential to be short ranged acting over a length scale of water molecule and can be offset by the solute-water attractive interactions.

Indeed, the attractive methane-water interaction increases the contact water density at methane surface by 10% compared to its equivalent hard sphere \[81\]. In the simulation study of hydration of methane clusters, when methane-water attractions were turned off the water dewets the cluster as predicted by scaled particle theory \[66, 67, 68\], but rewets the cluster when the attractive potential was turned
Similarly, when the neopentane-water attraction was turned off, the solvent separated minimum disappears in the potential of mean force between two neopentane molecules \[\text{[78]}\], indicating a dewetting like transition. The critical distance of separation of two oblate ellipsoidal surfaces \[\text{[83, 84]}\] or flat plates \[\text{[85]}\], for the onset of dewetting, is shorter for attractive hydrophobic surfaces than the purely repulsive surfaces. Even a minute reduction in attraction between the wall and water interactions can dramatically affect the water conduction through a nanotube, leading to a sharp two state transition between the empty and the water filled state \[\text{[86]}\].

The extent of attractive interactions of a constituent atoms of the surface with water determines the hydration of the surface. For example, liquid water confined between hydrophobic surfaces becomes unstable and a vapor phase is formed. Whereas water confined between hydrophilic surfaces at the similar condition remains stable \[\text{[87]}\]. This prompted many research groups to investigate the wetting behavior of parallel plates with different patterns of hydrophobic and hydrophilic patches \[\text{[88, 89, 90]}\]. From extensive simulation studies, it was found that presence of hydrophilic patches either delays or totally prevents the dewetting of the inter-plate region. It was observed that walls with substantial polarity, \(i.e.,\) surface dipole moment of \(\sim 40\%\) of dipole moment of water, behave hydrophobic macroscopically and dewets forming a vapor phase \[\text{[91, 92, 90]}\].

The drying of the interstitial region between large hydrophobic surfaces are successfully predicted by a simple macroscopic thermodynamic model based on Young’s law \[\text{[84]}\]. Similarly, Kelvin equation describing capillary condensation and evaporation, qualitatively explains the stability of vapor state in narrow hydrophobic nano
pores (R ≤ 0.45 nm), liquid filled state in large pores (R ≥ 0.65 nm) and the liquid-vapor oscillations in the intermediate pore sizes [93]. Such macroscopic theory, however, cannot explain the dewetting observed on a single surface, as it occurs over molecular distances.

An unified theory on molecular and macroscopic hydrophobic hydration by Lum et al., [75] successfully predicts the characteristic thermodynamic anomalies exhibited by hydration of small hydrophobic molecules, the onset of liquid-vapor phase transition in confined geometries and the resultant long-range hydrophobic forces between surfaces. Recently, Willard and Chandler [94] have proposed a coarse-grain model describing the important characteristics of the wall-water interface, including the spatial dependence of the interface width and large density fluctuations of water at interface, for flat heterogeneous surfaces. The model, also, captures the effect of attractive interactions in pulling the interface closer to the surface.

Surface roughness and curvature are other important factors influencing the wetting behavior of the surfaces. Lee and Rossky [95] found a pronounced increase in water density on an atomically rough hydrophobic surface, attributed to a local higher attractive potential between a water molecule and LJ molecules comprising the surface. We have also observed that spherically rough surface are preferentially hydrated and water molecules are found to be thermodynamically and kinetically bound to the surface [96]. However, the structural and thermodynamic properties of water in spherical cavities, negative curvature, is similar to that observed in gas phase [97, 98]. Though the dispersion attraction decreases the interface width, an increased water density at the wall as noticed for flat or convex surfaces was not observed.
The wetting behavior of surfaces and stability of water in confinement depends on the curvature of the surface and wall-water interactions. Many simulation studies have investigated the effect of interaction potential on the hydration of large surfaces, however, the effect of curvature have not been studied extensively. In this work, we systematically study the effect of curvature – concave and convex – on a single system, i.e., a hydrophobic sinusoidal wall, through molecular dynamics simulations. We also examine the extent of influence of wall-water interaction potential on the wetting and dewetting characteristics of concave and convex structures. We noticed that water wets both the concave and convex regions of the wall at full LJ potential for wall-water interactions. With decrease in wall-water attractions, the concave pockets dewet readily than the convex surfaces. The observed dewetting is only weak as the free energy of a water molecule in the dewetted interface is lower than the free energy of a water molecule at vapor-liquid interface.

3.2 Methods

3.2.1 Molecular Simulations

All molecular dynamic (MD) simulations were performed at constant NVT using GROMACS 4.0.3 [99]. The temperature was held constant at 300 K using Nosé-Hoover thermostat [100] [101] with a time constant of 0.1 ps. The sinusoidal wall depicted in Fig. 3.1 is made of 6144 close-packed Lennard-Jones (LJ) spheres ($\epsilon_{ss} = 0.09$ kcal/mol and $\sigma_{ss} = 3.27$ Å) with dimensions given in the figure caption. The wall was placed in the center of the simulation box and solvated by 14209 SPC/E water molecules [102]. The positions of the LJ spheres were held fixed during the simulation. Simulations were carried for different interaction potentials between water and the
Figure 3.1: Sinusoidal Lennard-Jones (LJ) wall comprised of 6144 closed-packed LJ atoms separated by 2.5 Å, the center-to-center distance between adjacent atoms. The wall extends over a width of 80 Å in the x-direction, and a length of 80 Å in the y-direction. The well depth is 20 Å in the z-direction, measured from the center of the LJ atom in the bottom center of the well to the center of the LJ atom at the top. The atoms comprising the convex and the concave surfaces of the wall are highlighted in green and yellow, respectively.

LJ spheres by scaling the attractive interactions from $\lambda = 1.0$ to $\lambda = 0.3$ using the WCA decomposition of LJ pair potential [103]:

$$\phi_{sw}(r; \lambda) = \phi_{rep}(r) + \lambda \phi_{att}(r)$$  \hspace{1cm} (3.1)$$

where,

$$\phi_{rep}(r) = \begin{cases} 
4\epsilon_{sw} \left[ \left( \frac{\sigma_{sw}}{r} \right)^{12} - \left( \frac{\sigma_{sw}}{r} \right)^{6} \right] + \epsilon_{sw} & \text{if } r < 2^{1/6}\sigma_{sw} \\
0 & \text{if } r \geq 2^{1/6}\sigma_{sw}
\end{cases}$$

$$\phi_{att}(r) = \begin{cases} 
-\epsilon_{sw} \left[ \left( \frac{\sigma_{sw}}{r} \right)^{12} - \left( \frac{\sigma_{sw}}{r} \right)^{6} \right] & \text{if } r < 2^{1/6}\sigma_{sw} \\
4\epsilon_{sw} \left[ \left( \frac{\sigma_{sw}}{r} \right)^{12} - \left( \frac{\sigma_{sw}}{r} \right)^{6} \right] & \text{if } r \geq 2^{1/6}\sigma_{sw}
\end{cases}$$

The LJ parameters for wall-water interactions are obtained by applying Lorentz-Berthelot mixing rules: $\epsilon_{sw} = \sqrt{\epsilon_{ss}\epsilon_{ww}}$ and $\sigma_{sw} = (\sigma_{ss} + \sigma_{sw})/2$.  

32
Table 3.1: Simulation box length in the z-direction (in Å) for simulations at different LJ pair potentials defined by the scaling parameter λ (Eq. 3.1).

<table>
<thead>
<tr>
<th>λ</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>L_z</td>
<td>96.0</td>
<td>95.0</td>
<td>94.0</td>
<td>92.5</td>
<td>88.0</td>
<td>86.2</td>
<td>85.4</td>
<td>85.0</td>
</tr>
</tbody>
</table>

Pressure is an important thermodynamic parameter influencing the hydration of surfaces [97, 88, 89, 104]. To insure that all simulations were run at a pressure of \( \sim 1 \) bar or slightly higher, we carried out a series of 250-ps NVT simulations, and iteratively adjusted the simulation box volume in the z-direction to obtain the bulk density of liquid water at 300 K and 1 bar far from the wall (\( z \geq 35 \) Å from the box center). The adjusted simulation box lengths in as a function of λ are given in Table. 3.1.

Electrostatic interactions were computed using the particle-mesh Ewald method with a real space cut-off of 14 Å. Non-bonded LJ interactions were truncated at 13 Å. The geometry of water molecules was maintained by constraining water oxygen-hydrogen bond lengths using SETTLE algorithm [105]. Each system was equilibrated for at least 500 ps. Configurations were saved every 0.1 ps over the 2-ns production run for all λ, except 1.0 and 0.5. Production runs of 10 ns were carried out for \( \lambda = 1.0 \) and \( \lambda = 0.5 \).

3.2.2 Analysis

The potential of mean force, \( W(r) \), to bring a water molecule from bulk solution to a distance \( r \) from the concave or convex surface of the LJ wall was computed from
the proximal distribution of water [82] around the wall atoms of interest, where \( r \) is the minimum distance to any atom that defines the concave or convex surface of the wall (Fig. 3.1). The accessible volume for waters proximal to these concave and convex surfaces is calculated by constructing a network of 0.05 Å cubic grid points, and finding those grid points that are within a distance \( r \) of the surface of interest and do not overlap with any wall atoms. The accessible volume is then given by the number of grid points multiplied by the cubic volume of each grid point [23, 96]. The accessible volume as the function of the well height, \( z \), was similarly evaluated. A bin size of 0.1 Å was used for all calculations of water distribution functions.

We note that because the concave and convex surfaces have different curvature, the proximal volume for the convex surface increases with increasing \( r \), while the proximal volume for the concave surface decreases with increasing \( r \). The latter limits the determination of \( W(r) \) for the concave surface to a maximum distance of \( r = 8 \) Å.

The excess chemical potential of water, \( \mu_{w}^{ex} \), was evaluated as using the quasi-chemical theory organization of the potential distribution theorem [29, 106],

\[
\beta \mu_{w}^{ex} = \ln x_0 - \ln p_0 - \ln \left\langle e^{-\beta \epsilon} | n = 0 \right\rangle_0 .
\]  

\( p_0 \) is the probability of finding a defined observation volume to be empty (cavity) and the term \(-kT \ln p_0\) is the free energy of the cavity formation. The rightmost term in Eq. 3.2 accounts for the energy of interaction of a water molecule, inserted into the center of the cavity, with the rest of the system consisting of water and wall atoms. The double brackets with subscript 0 specifies that the interaction energy is to be averaged over all the possible configurations of the inserted water molecule and the system, uncoupled from one another. The chemical association term \( kT \ln x_0 \) is the
free energy gain when other water molecules interact with the inserted water molecule. This is assessed by calculating the probability, $x_0$, of finding a water molecule without any neighboring water molecules within the defined observation volume.

The observation volume is defined by conditioning radii of exclusion of 3.3 Å for water oxygens and of 3.5 Å for wall atoms. Cavities are found by constructing a network of 1 Å cubic grid points, and identifying those points that have no water oxygens within 3.3 Å or wall atoms within 3.5 Å of each grid point. The probability, $p_0$, is obtained as the ratio of the number of these grid points to the total number of grid points available for water occupancy. $x_0$ is computed as the probability of finding a water molecule without any water molecules within a radial distance of 3.3 Å and any wall atoms within a radial distance of 3.5 Å. The interaction energy is calculated by randomly inserting a water molecule in 25 different orientations at each of these grid points. Electrostatic interactions were computed using generalized reaction field [107, 108, 109, 110] with a cut-off radius of 12 Å. LJ interactions were also truncated at 12 Å. The interaction energy contribution was obtained by numerically integrating the Boltzmann weighted probability distribution of interaction energies over the finite range of energies sampled in the simulation,

$$\langle\langle e^{-\beta\epsilon}|n=0\rangle\rangle_0 = \int_{-\infty}^{\epsilon_{\text{max}}} P^0(\epsilon|n=0)e^{-\beta\epsilon}d\epsilon$$

$$\approx \int_{\epsilon_{\text{min}}}^{\epsilon_{\text{max}}} P^0(\epsilon|n=0)e^{-\beta\epsilon}d\epsilon.$$  

3.3 Results and Discussion

Representative snapshots taken from MD simulations at different $\lambda$-scaling of wall-water attractions (Eq. 3.1) are shown in Fig. 3.2. The corresponding normalized water
Figure 3.2: Snapshots from MD simulations of the hydration of the LJ wall (Fig. 3.1) as attractive interactions between the wall atoms and water molecules are scaled from $\lambda = 1.0$ to $\lambda = 0.3$ (Eq. 3.1). Top row (left to right): $\lambda = 1.0, 0.9, 0.8$ and 0.7. Bottom row (left to right): $\lambda = 0.6, 0.5, 0.4$ and 0.3.
densities as the function of well height are shown in Fig. 3.3. For $\lambda = 1.0$ (full wall-water attractions), water completely fills the four wells of the LJ wall. There is also no noticeable difference in water densities proximal to the convex and the concave surfaces of the wall. The local density of the water in the well is on average about 10% greater than the bulk density of water. As $\lambda$ is reduced to $\sim 0.7$, however, the water density near the concave surface at the bottom of the well decreases, while the water density outside the well near the convex surface of the wall is unperturbed. The vapor-liquid interface, defined as the value of $z$ at which the density is half that of bulk water, is located at $\sim 7.5$ Å for $\lambda = 0.7$. As $\lambda$ is decreased below 0.6, this interface abruptly jumps to $z \geq 14$ Å, and then gradually shifts to $z \sim 17$ Å as $\lambda$ is decreased further to 0.3. The local water density outside the wells near the convex surface of the wall is not significantly affected as $\lambda$ is reduced from 1.0 to 0.3.

The bottom panel of Fig. 3.3 shows water density profiles across the width of the well at a height of $13 \text{ Å} < z < 16 \text{ Å}$ for different $\lambda$-scaling of attractive interactions between the wall atoms and water molecules. As $\lambda$ decreases from 1.0 to 0.3, the peak at $x \sim 9$ Å decreases, while the broad plateau between 0 and 4 Å increases, indicating that water in the well moves away from the confining surfaces and towards the center of the well.

The potential of mean force, $W(r)$, is shown in Fig. 3.4 as attractive interactions between the wall atoms and water molecules are reduced from $\lambda = 1.0 \rightarrow 0.7$ for water proximal to the concave surface, and $\lambda = 1.0 \rightarrow 0.3$ for water proximal to the convex surface. For full wall-water attractions, $W(r)$ has two minima, a contact minimum at $3.4$ Å and a solvent-separated minimum at $6.5$ Å, with the contact minimum the more stable minimum. The contact minimum corresponding to water near the convex
Figure 3.3: Top panel: Normalized water density as the function of the well height, \( z \) as attractive interactions between the wall atoms and water molecules are scaled from \( \lambda = 1.0 \) to \( \lambda = 0.3 \) (Eq. 3.1). Bottom panel: Water density profile across the width of the concave well at a height of \( 13 \, \text{Å} < z < 16 \, \text{Å} \) for the same range of \( \lambda \)-scaling of the wall-water attractive interactions.
Figure 3.4: Potential of mean force of bringing a water molecule from bulk solution to a distance, $r$, from any atom in the concave (top panel) or convex (bottom panel) surfaces of the LJ wall (Fig. 3.1) as attractive interactions between the wall atoms and water molecules are scaled from $\lambda = 1.0$ to $\lambda = 0.7$ (top panel) or $\lambda = 1.0$ to $\lambda = 0.3$ (bottom panel).
surface is $\sim 0.3$ kT lower than that for water near the concave surface, revealing the curvature dependence of hydration of these hydrophobic surfaces.

In both cases, decreasing wall-water attractions first destabilizes the contact minimum and then the solvent-separated minimum, although the effect is much more dramatic for water proximal to the concave surface. For the concave surface, the solvent-separated minimum is slightly more stable at $\lambda = 0.9$, and both minima vanish as wall-water attractions are reduced to $\lambda = 0.8$, whereas a shallow contact minimum is still apparent at $\lambda = 0.4$ for water proximal to the convex surface, and the solvent-separated minimum becomes the more stable minimum only at $\lambda = 0.3$.

The observation that large hydrophobic surfaces dewet in the absence of wall-water attractions has been well established in earlier simulation studies [82, 83, 84, 85]. The influence of curvature on hydrophobic hydration has also been observed, for example, in simulations of the hydration of an oblate ellipsoidal solute compared to a spherical solute of same volume [111]. Hydrophobic hydration for convex and concave geometries has been investigated by several research groups [83, 97, 86, 98, 93]. However, direct comparisons of the curvature dependence of hydration is limited by differences in the LJ parameters used in the different studies. The curvature dependence observed here is also predicted based on the cavity expulsion potential derived for spherical LJ solutes as a function of solute size [80].

Fig. 3.5 shows the average water occupancy in a spherical observation volume of radius 3.3 Å at two locations: 3.4 Å from the bottom center of the well, and 3.4 Å above the center of the convex surface as a function of $\lambda$-scaling of wall-water attractions. As the wall-water attractions are reduced, the average water occupancy decreases at both locations. However, the effect is more pronounced at the bottom of
the well, near the concave surface, than it is outside the well, near convex surface, as expected from the water potentials of mean force shown in Fig. 3. For comparison, the average occupancy in a spherical observation volume of radius 3.3 Å in bulk water \((z > 25 \, \text{Å})\) is shown as a horizontal dashed line at \(\langle N \rangle = 5.05\).

Water densities in the bottom of the well are also expected to be strongly correlated along the length of the well in the y-direction. This correlation is shown in Fig. 3.6 by plotting

\[
C_{12} = \frac{\langle N_1 N_2 \rangle - \langle N_1 \rangle^2}{\langle N_1^2 \rangle - \langle N_1 \rangle^2},
\]

where \(N_1\) and \(N_2\) are the water occupancies in two identical observation spheres of radius 3.3 Å separated by a certain center-to-center distance in the y-direction. Fig. 3.6 compares the y-dependence of \(C_{12}\) as a function of \(\lambda\)-scaling for two spheres located at \(z = 3.4 \, \text{Å}\) above the bottom center of the well, and at \(z = 3.4 \, \text{Å}\) above the center of the convex surface outside the well. \(C_{12}\) decays exponentially in both cases, but the correlations are clearly much longer-ranged and more sensitive to the wall-water attractive interactions at the bottom of the LJ well. The dramatic change in \(C_{12}\) for water occupancies in the bottom of the well in response to \(\lambda = 0.9 \rightarrow 0.8\) corresponds to the loss of both the contact minimum and solvent-separated minimum in the potential of mean force in Fig. 3, and suggests that the confinement of water in the well enhances intermolecular interactions with adjacent water molecules located along the length of the well. Conversely, water molecules hydrating the convex surface outside the well interacts with adjacent water molecules laterally, as well as long the length of the well, and as a consequence, these long-ranged correlations are not observed. The behavior is similar to that observed for water conduction through the hydrophobic channel of a carbon nanotube [86].
Figure 3.5: Average number of water molecules in spherical observation volumes of radius 3.3 Å located 3.4 Å above the bottom center of the concave surface (solid circles) and 3.4 Å above the center of the convex surface (open circles) of the LJ wall as a function of $\lambda$ scaling of the wall-water attractive interactions. The bulk water occupancy number in an identical observation volume located 30 Å above the convex surface is shown as a horizontal dashed line. Lines connecting the points show the trends in the data.
Figure 3.6: Correlation of water occupancies in two spherical observation volumes of 3.3 Å radius located 3.4 Å above the bottom center of the concave surface (red lines) and 3.4 Å above the center of the convex surface (black lines) as a function of the center-to-center distance separating the observation volumes in the y-direction. The lines correspond to different $\lambda$ scaling of wall-water attractive interactions, with the arrows showing the direction of decreasing $\lambda$. For the red lines, $\lambda = 1.0, 0.9, 0.8$ and 0.7. For the black lines, all values of $\lambda$ between 1.0 and 0.3.
Figure 3.7: The variation of excess chemical potential of water (solid triangles) and its corresponding quasi-chemical contributions: $kT \ln x_0$ - open circles; $-kT \ln p_0$ - solid circles; outer shell contributions - open triangles - as the function of well height when water completely wets and partially dewets the wall, i.e., when the wall-water attraction is completely turned on, $\lambda = 1.0$, – top panel – and when the attraction is partially (50%) turned off, $\lambda = 0.5$ – bottom panel. The dashed line is the experimental value of excess free energy of water, -6.32 kcal/mol [1].
Figure 3.8: Logarithm of conditional probability distribution of binding energies, conditioned such that there is no water within 3.3 Å spherical observation volume about the central water oxygen, of waters in the concave well when the wall-water attraction is completely turned on, $\lambda = 1.0$, top panel – and when the attraction is partially (50%) turned off, $\lambda = 0.5$ – bottom panel – at different well heights, $z$: $4 < z < 5$ Å - black circles; $9 < z < 10$ Å - red circles; $14 < z < 15$ Å - green circles; $z = 19 < z < 20$ Å - blue circles; $z = 24 < z < 25$ Å - violet circles; $29 < z < 30$ Å - dark green circles; $34 < z < 35$ Å - orange circles. The lines are the gaussian fit. The consecutive energy distributions are shifted by 4 units for visual clarity.
Fig. 3.7 shows the excess chemical potential of water and the quasi-chemical contributions to \( \mu^e_w \) (Eq. 3.2) as a function of \( z \), for \( \lambda = 1.0 \) and \( \lambda = 0.5 \). For \( \lambda = 1.0 \), \( \mu^e_w \) for water in the well is close to that in bulk solution \( (z > 25) \); \( \mu^e_w = -6.7 \) kcal/mol averaged over all \( z > 4 \) Å. For \( \lambda = 0.5 \), \( \mu^e_w \) gradually increases starting at \( z \sim 15 \) Å from the bulk solution value to approximately \( \mu^e_w = -2.0 \) kcal/mol at \( z = 7 \) Å. The standard deviation for \( \mu^e_w \) as a function of \( z \) was determined to be \( \sim 0.5 \) kcal/mol on averaged, with the larger standard deviations calculated for the bulk solution \( (z > 25) \).

The quasi-chemical contributions, \(-kT \ln p_0\) and \(kT \ln x_0\), essentially compensate one another in bulk solution \( (z > 25) \), as expected for the conditioning radii used [112]. Although the outer shell contribution is smaller, it cannot be neglected. The probability distributions of interaction energies for \( \lambda = 1.0 \) and \( \lambda = 0.5 \) as a function of well depth, \( z \), are shown in Fig. 3.8. The distributions are satisfactorily gaussian in bulk solution, as expected [113], but are distinctly non-gaussian in the well \( (z < 20 \) Å). Non-gaussian binding energy distributions were also observed in the water vapor-liquid interface [114].

For \( \lambda = 1.0 \), the work required to create a cavity in water is most sensitive to the confinement of water in the well. This quasi-chemical contribution decreases with decreasing \( z \), as water loses many of its hydrogen bonding partners. For \( \lambda = 0.5 \), all the quasi-chemical contributions vary significantly below the water vapor-liquid interface at \( z \sim 15 \) Å as the water density decreases. The overall change in \( \mu^e_w \) is approximately \( +4 \) kcal/mol between bulk solution and the bottom of the well, which is less than the change in \( \mu^e_w \) across the water vapor-liquid interface: \( 6.5 \) kcal/mol [114],
which implies weak dewetting of the concave hydrophobic surface at the bottom of the well \[80\].

### 3.4 Conclusions

We investigated the hydration of a hydrophobic sinusoidal wall, and specifically, the influence of wall-water attractive interactions and surface curvature on hydration. As wall-water attractive interactions are reduced, dewetting of wall was observed. The wetting behavior of the concave surface of the wall was more sensitive to the wall-water attractions compared to the convex surface. In addition, we observed long-ranged correlations in water densities along the length of the wall above the concave surface at the bottom well.

A quasi-chemical theory analysis of the excess chemical potential of water as a function of well depth determined that the work required to form a cavity in water decreases with the confinement of water in the well. In addition, the probability distribution of water interaction energies are gaussian in bulk solution, but becomes distinctly non-gaussian in the well, indicating strong correlations in water intermolecular interactions in the well. Finally, our analysis reveals only a weak dewetting of the concave hydrophobic surface, as the free energy for transferring a water molecule from bulk solution to near hydrophobic wall is lower than the free energy for transferring a water molecule from bulk solution to the vapor phase.
CHAPTER 4

Aqueous Partial Molar Volumes of N-Acetyl Amino Acid Amides from Simulation and Group Contributions

4.1 Introduction

Conformational stability and dynamics of proteins in aqueous solutions can be identified and characterized using protein volumetric properties such as partial molar volume (PMV) and its pressure and temperature derivatives and their fluctuations [115]. For example, the difference in PMV with increasing pressure for different conformational states of a protein determines the population of the stable conformations and eventually predicts protein denaturation at high pressures. The experimental PMVs of many globular proteins fall within a narrow range – 0.70 – 0.74 cm$^3$/g [116, 117] due to the similarity in the overall amino acid composition of protein surfaces: 57% non-polar, 24% polar and 19% charged (within 5% root mean square variations) [118], despite a wide variation in shapes and sizes. These similarities have prompted the proposition of empirical group contribution schemes for protein PMVs based on its amino acid composition [119, 120, 121, 122, 117, 123]. Contributions from the protein backbone and amino acid side-chains are considered to be purely additive, with corrections for charges on the N– and C– terminals, and side-chain ionization. The backbone contribution per amino acid residue is usually taken as the
PMV of a central glycyl group in a short poly-glycine peptide, while the difference in the PMV for other amino acids with respect to glycine provides side-chain contributions. PMVs are measured experimentally for zwitter ions [119, 124, 125, 126], GXG tripeptides where X is the amino acid of interest end capped by glycine at either ends [121], polypeptides end capped by glycine at either ends [122], and N-acetyl amide derivatives of amino acids [127, 125, 128, 129, 130]. Empirical group additivity schemes concentrate on the protein chemistry neglecting topological heterogeneity. However, the geometry of the protein surface is found to be responsible for two-thirds of water structure in the first hydration layer, while the influence of amino acids chemistry through electrostatics and hydrogen bonding interactions, is responsible for only one-third of the observed water structure [131].

The PMV of a solute at infinite dilution can be obtained from the microscopic solvent structure, i.e., the molecular packing of solvent around solute molecule using Kirkwood-Buff (KB) theory [31]. Hirata and coworkers estimated PMVs of all 20 amino acids through KB theory by employing a one-dimensional reference interaction site model (1D-RISM) with the hypernetted chain approximation to determine solute-solvent and solvent-solvent direct correlation functions [132]. The estimations were further improved using a three-dimensional reference interaction site model (3D-RISM) for solute-sovlent correlation functions [133]. PMVs were significantly underestimated, however, especially those with bulky and aromatic side-chains, as the theoretical approach did not account for the conformational flexibility of the amino acids. The solute-solvent pair correlation function can be readily obtained from molecular simulations [33, 34, 35, 134]. Recently PMVs for molecular solutes have been determined from the proximal distributions of water around the constituent groups
comprising these solutes [34, 35, 134]. The proximal distributions of water around each identical groups were found to be identical across the solutes considered, indicating that the water structure is only locally sensitive to the chemical details of the individual groups. The PMVs for 28 organic solutes, comprising a wide range of functional groups were calculated from the proximal distribution of water around 11 chemical groups [134]. Local water distributions proximal to carbon, nitrogen or oxygens groups [131, 135] on protein surfaces have also been observed to be independent of the different proteins studied.

In this work, we investigate the hydration of the 20 naturally occurring amino acids by molecular dynamics simulation, and employ KB theory to compute its partial molar volume from the simulation data. In order to examine hydration without any influence of the charged amino and carboxy terminal, the end groups are capped by acetyl and amide groups, respectively. Local water densities around an amino acid are computed from the proximal distributions of water around each heavy atom. The PMVs calculated from proximal distributions are only slightly (~5 %) and consistently over-predicted. PMVs of side-chains of polar and non-polar amino acids when obtained as the difference of PMV of amino acid and PMV of glycine closely matches the experimental data. The calculated PMVs for charged amino acids are significantly lower than the corresponding experimental values. PMVs of amino acids evaluated by the group contribution scheme, defined by the averaged proximal water distribution for identical groups across all amino acids, is only qualitative, as cooperative hydration [136] due to the chemical nature of the adjacent groups cannot be neglected.
4.2 Theory

Kirkwood-Buff (KB) theory \[31\] relates macroscopic thermodynamic quantities like PMV, chemical potential to the microscopic fluctuations in a open system. The PMV of a solute, at infinite dilution, is given by,

\[
\overline{V}_s^\infty = kT\kappa_0 - 4\pi \int_0^\infty (g_{sw}(r|\mu VT) - 1)r^2 dr .
\] (4.1)

The first term on the right side of this equation corresponds to the ideal contribution due to inherent solvent fluctuation, where, \(kT\) is the thermal energy – a product of Boltzmann constant \(k\), and temperature \(T\), and \(\kappa_0\) is isothermal compressibility of the solvent. For TIP3P water \(\kappa_0 = 4.2 \times 10^{-5} \text{ atm}^{-1}\) \[134\]. While the second term on the right side of this equation is the KB integral or the excess part of the PMV owing to the solute-solvent interactions. The \(g_{sw}(r|\mu VT)\) is the radial distribution function (rdf) of water (\(w\)) around the solute (\(s\)) defined in a open or grand canonical system.

The application of KB theory in a canonical system requires a thermodynamic transformation to account for the difference in asymptotic solvent bulk densities normalizing the solute-solvent rdfs. The KB expression \[33, 34\] for calculating PMV from a canonical simulation is

\[
\overline{V}_s^\infty = kT\kappa_0 + \frac{4\pi \int_0^{R_c} r^2 dr - 4\pi \rho_{w,b}^{-1} \int_0^{R_c} \rho_{sw}(r)r^2 dr}{1 - 4\pi N_T^{-1} \int_0^{R_c} \rho_{sw}(r)r^2 dr} .
\] (4.2)

Here, \(\rho_{sw}(r)\) is the local density of solvent around the solute obtained from a simulation with fixed number of solute and solvent molecules, \(\rho_{w,b}\) is the bulk solvent density and \(N_T\) is the total number of solvent molecules in the simulation, and \(R_c\) is the radial distance from the solute center, up to which the solvent density is correlated with the presence of the solute.
By taking appropriate derivatives of Eqs. 4.1 and 4.2, an effective grand canonical radial distribution [134], \( g_{sw}(R_c) \), is defined for a closed system,

\[
g_{sw}(R_c) = 1 - \frac{1 - \rho_{w,b} \rho_{sw}(R_c)}{1 - 4 \pi N_T^{-1} \int_0^{R_c} \rho_{sw}(r)r^2 dr} \frac{\rho_{sw}(R_c) \left( 4 \pi \int_0^{R_c} r^2 dr - 4 \pi \rho_{w,b} \int_0^{R_c} \rho_{sw}(r)r^2 dr \right)}{N_T \left( 1 - 4 \pi N_T^{-1} \int_0^{R_c} \rho_{sw}(r)r^2 dr \right)^2},
\]

such that

\[
\bar{V}_s^\infty = k T \kappa_0 - 4 \pi \int_0^{R_c} (g_{sw}(r) - 1)r^2 dr;
\]

and the solvent density for \( r < R_c \) equals the solvent density for \( r > R_c \) which is equivalent to the corresponding excess chemical potential of the solvent are equal.

Knowing the water proximal rdfs [137] for the constituent atoms of the solute, say \( n \) heavy atoms, the PMV of the solute can be obtained [134]:

\[
\bar{V}_s^\infty(R_c) = k T \kappa_0 - \sum_{\alpha=1}^{n} \int_0^{R_c} \left( g_{sw,\alpha}(r) - 1 \right) \frac{\partial V_{\alpha}(r)}{\partial r} dr,
\]

where each proximal rdf, \( g_{sw,\alpha}(r) \), is integrated over its corresponding proximal volume, \( V_{\alpha} \). Proximal rdfs can be determined by partitioning space around the solute into proximal subdomains centered on each heavy atom [134, 137, 138]. For a molecular solute defined by \( n \) sites, the proximal solute site \( \alpha \) at \( \vec{r}_\alpha \) for a water molecule at \( \vec{r} \) is determined by

\[
|\vec{r} - \vec{r}_\alpha| - \sigma_\alpha = \min_{\beta=1,\ldots,n} |\vec{r} - \vec{r}_\beta| - \sigma_\beta.
\]

Size disparities among the solute sites are taken into account in Eq. 4.6 by the solvent excluded-volume radius, \( \sigma_\beta \). This radius for a solute site is computed as the distance at which the corresponding proximal rdf first equals unity.

Although proximal rdfs are readily obtained from a simulation, the long range oscillations in rdfs get magnified in the integrals (Eqs. 4.4 and 4.5) leading to poor
convergence of PMVs. The convergence can be improved by smearing the rdfs assuming multiple solvent reference centers for water – all points within a spherical radius of 2.3 Å about its oxygen center is considered to be an equiprobable location of the water molecule [34, 35, 134] – instead of single reference point, e.g., water oxygen.

4.3 Simulation Details

All molecular dynamics simulations were carried out using NAMD 2.6 [58] and CHARMM-27 force field [139] with an integration time step of 2 fs. Each end-capped amino acid was solvated by 3000 TIP3P [140] water molecules. The systems were initially relaxed by executing 20,000 steps of conjugate gradient energy minimization, and then equilibrated for 100 ps at constant NVT. To determine the equilibrium volume of a system at 300 K and 1 bar, the system was equilibrated at constant NPT for 500 ps. The average box volume computed over the last 300 ps of this simulation, was used in the subsequent constant NVT simulation. Following a short final equilibration for 500 ps, configurations were saved for analysis every 0.1 ps over a production run of 5 ns simulation.

Temperature was held constant at 300 K during the simulation by applying Langevin dynamics to all heavy atoms using a damping coefficient of 1 ps$^{-1}$. Equilibration at a constant pressure of 1 bar was carried using a Nosé-Hoover Langevin piston [141, 142] with a period of 200 fs and a decay of 100 fs. Periodic boundary conditions were imposed, and the particle-mesh Ewald method [143] with a real-space cutoff of 12 Å was used in computing electrostatic interactions. Lennard-Jones interactions were smoothly switched to zero at 12 Å starting from 10 Å. All hydrogen atom positions were constrained using the SHAKE algorithm [60].
4.4 Results and Discussion

The constituent groups of the 20 naturally occurring amino acids have been classified into the 11 chemical groups listed in Table 4.1, similar to the simulation study on organic solutes [134]. The methyl/methylene groups are sub-classified depending on the adjacent bonded atom, all nitrogen groups are classified as a single group. The nitrogen atom mainly occurs as –NH$_2$ and –NH$^-$ in aliphatic carbon chains and –NH$^-$ in the cyclic ring of proline and in aromatic rings of tryptophan and histidine. Hydration of the two predominant nitrogen groups showed little variation. Even their vdW radii varied only slightly compared to variations in methyl/methylene groups [134].

Proximal regions for these 11 groups were computed by Monte Carlo technique [137] as the fraction of spherical volume about the central atom of the group that is proximal to the group at a given distance. The proximal volume around the group was calculated as an average over different amino acids conformations. Thus, proximal water rdfs about the groups are determined. The solvent excluded volume radius for each group are given in Table 4.1 and differ slightly from that reported in [134] as the force fields used in these two studies are different.

Fig. 4.1 shows the effective grand canonical radial distribution function for water around N-acetyl-alanine-amide determined from the proximal radial distribution functions of water about its constituent molecular groups –

$$g_{sw}(r) = \frac{\sum_{\alpha=1}^{11} g_{sw,\alpha}(r) \frac{\partial V_{\alpha}(r)}{\partial r}}{\sum_{\alpha=1}^{11} \frac{\partial V_{\alpha}(r)}{\partial r}} .$$ (4.7)

The figure also shows the PMV obtained from the proximal distribution functions. The oscillations in the calculated PMV are an attribute of long range oscillations in
Table 4.1: Classification of constituent groups of the amino acids with their solvent excluded volume radii

<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
<th>$\sigma$ (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-N-$</td>
<td>amine and other nitrogen groups</td>
<td>2.74</td>
</tr>
<tr>
<td>O=(C)</td>
<td>carbonyl oxygen group</td>
<td>2.52</td>
</tr>
<tr>
<td>OH</td>
<td>hydroxyl oxygen group</td>
<td>2.61</td>
</tr>
<tr>
<td>$-S-$</td>
<td>sulphur group</td>
<td>3.15</td>
</tr>
<tr>
<td>C=(O)</td>
<td>carbonyl carbon group</td>
<td>3.09</td>
</tr>
<tr>
<td>ACH or AC−(C)</td>
<td>aromatic carbon group including those connected to a carbon group</td>
<td>3.15</td>
</tr>
<tr>
<td>AC−(O/N)</td>
<td>aromatic carbon connected to a polar oxygen/nitrogen group</td>
<td>2.96</td>
</tr>
<tr>
<td>CH</td>
<td>methylene (CH) group</td>
<td>3.31</td>
</tr>
<tr>
<td>CH$_{2/3}$−(O/N)</td>
<td>methyl or methylene group connected to a polar oxygen/nitrogen group</td>
<td>3.24</td>
</tr>
<tr>
<td>CH$_2$−(C)</td>
<td>methylene group connected to a carbon group</td>
<td>3.29</td>
</tr>
<tr>
<td>CH$_3$−(C)</td>
<td>methyl group connected to a carbon group</td>
<td>3.26</td>
</tr>
</tbody>
</table>
Figure 4.1: Top panel: The effective grand canonical radial distribution function for water, smeared (solid line) and unsmeared (dashed line), around N-acetyl-alanine-amide determined from the proximal radial distribution of water around its constituent groups. Bottom panel: PMV of N-acetyl-alanine-amide computed from Eq. 4.5.
Table 4.2: PMV of the N-acetyl amino acid amides at infinite dilution in water at 300 K, determined from simulations (\( \overline{V}_\text{Sim}^\infty \)) and group contributions (\( \overline{V}_\text{GC}^\infty \)) using KB theory. PMV contributions from side-chain (\( \overline{V}_\text{SC}^\infty \)) and backbone (\( \overline{V}_\text{B}^\infty \)), including C\( \alpha \) carbon atom, of amino acids are also listed.

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>( \overline{V}_\text{SC}^\infty )</th>
<th>( \overline{V}_\text{B}^\infty )</th>
<th>( \overline{V}_\text{Sim}^\infty )</th>
<th>( \overline{V}_\text{GC}^\infty )</th>
<th>( \overline{V}_\text{Exp}^\infty )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>37.7 ± 0.5</td>
<td>86.0 ± 0.5</td>
<td>113.6 ± 0.9</td>
<td>102.7 ± 0.5</td>
<td>108.0</td>
</tr>
<tr>
<td>Val</td>
<td>68.6 ± 1.5</td>
<td>82.0 ± 2.4</td>
<td>148.0 ± 0.7</td>
<td>133.8 ± 0.4</td>
<td>139.0</td>
</tr>
<tr>
<td>Leu</td>
<td>84.9 ± 2.0</td>
<td>81.1 ± 2.7</td>
<td>166.1 ± 0.6</td>
<td>153.1 ± 0.3</td>
<td>155.9</td>
</tr>
<tr>
<td>Ile</td>
<td>84.6 ± 1.4</td>
<td>80.3 ± 2.4</td>
<td>162.5 ± 0.8</td>
<td>150.8 ± 0.4</td>
<td>153.9</td>
</tr>
<tr>
<td>Met</td>
<td>84.0 ± 2.4</td>
<td>79.1 ± 2.2</td>
<td>161.5 ± 0.8</td>
<td>156.9 ± 0.4</td>
<td>153.7</td>
</tr>
<tr>
<td>Pro</td>
<td>62.9 ± 2.0</td>
<td>76.8 ± 1.2</td>
<td>136.3 ± 1.3</td>
<td>119.0 ± 1.0</td>
<td>126.4</td>
</tr>
<tr>
<td>Gly</td>
<td>20.1 ± 0.3</td>
<td>94.3 ± 0.6</td>
<td>94.3 ± 0.6</td>
<td>86.3 ± 0.7</td>
<td>90.8</td>
</tr>
<tr>
<td>Ser</td>
<td>38.9 ± 1.7</td>
<td>85.6 ± 1.9</td>
<td>115.6 ± 0.9</td>
<td>106.1 ± 0.5</td>
<td>108.2</td>
</tr>
<tr>
<td>Thr</td>
<td>53.2 ± 1.6</td>
<td>82.8 ± 1.7</td>
<td>129.9 ± 1.3</td>
<td>113.7 ± 0.8</td>
<td>124.6</td>
</tr>
<tr>
<td>Phe</td>
<td>104.0 ± 2.0</td>
<td>75.7 ± 3.4</td>
<td>180.3 ± 1.0</td>
<td>172.7 ± 0.8</td>
<td>170.5</td>
</tr>
<tr>
<td>Trp</td>
<td>122.7 ± 1.7</td>
<td>76.8 ± 4.0</td>
<td>201.6 ± 0.6</td>
<td>194.0 ± 0.7</td>
<td>193.1</td>
</tr>
<tr>
<td>Tyr</td>
<td>102.1 ± 2.9</td>
<td>82.9 ± 2.1</td>
<td>185.5 ± 1.1</td>
<td>173.5 ± 0.6</td>
<td>174.0</td>
</tr>
<tr>
<td>Gln</td>
<td>70.1 ± 2.3</td>
<td>80.1 ± 2.2</td>
<td>148.7 ± 0.7</td>
<td>134.4 ± 0.9</td>
<td>142.1</td>
</tr>
<tr>
<td>Asn</td>
<td>55.0 ± 1.7</td>
<td>85.2 ± 1.2</td>
<td>132.3 ± 0.9</td>
<td>119.2 ± 0.8</td>
<td>125.5</td>
</tr>
<tr>
<td>Cys</td>
<td>51.6 ± 1.6</td>
<td>83.9 ± 1.6</td>
<td>129.5 ± 0.4</td>
<td>119.1 ± 0.6</td>
<td>–</td>
</tr>
<tr>
<td>His</td>
<td>71.1 ± 1.7</td>
<td>83.9 ± 0.8</td>
<td>149.3 ± 1.0</td>
<td>141.5 ± 0.6</td>
<td>–</td>
</tr>
<tr>
<td>Asp</td>
<td>35.1 ± 1.5</td>
<td>83.4 ± 1.8</td>
<td>113.8 ± 0.5</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Glu</td>
<td>46.9 ± 2.0</td>
<td>83.4 ± 0.9</td>
<td>125.7 ± 1.0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Lys</td>
<td>76.0 ± 2.9</td>
<td>82.1 ± 3.4</td>
<td>155.7 ± 1.1</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Arg</td>
<td>84.9 ± 2.1</td>
<td>79.9 ± 2.4</td>
<td>166.2 ± 0.7</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

\(^a\)obtained from Ref. [125, 127, 128, 129, 130]
Figure 4.2: Comparison of the PMV of the N-acetyl amino acid amides excluding cystine, histidine, lysine, arginine, aspartic acid and glutamic acid, calculated from simulations with their corresponding experimental values.

We observe that the PMV computed from the smeared distribution converges for $R_c > 10 \text{ Å}$. The PMVs of all N-acetyl amino acid amides calculated as an average value over $R_c = 10–15 \text{ Å}$ are reported in Table 4.2 along with the experimental PMVs for some of these amino acids [127, 125, 128, 129, 130]. The reported error bars are one standard deviation from the average values.

The PMVs are consistently over-predicted by $\sim 5\%$ as shown in Fig. 4.2 suggesting a systematic error largely due to the deficiency the TIP3P water model in predicting volumetric properties of water. It is well known that water oxygen-oxygen rdf of TIP3P water is devoid of the typical secondary hydration peak. Also, PMV of TIP3P water at 25°C and 1 atm pressure is 2% greater than the measured PMV of water [134]. The estimation of PMVs from water proximal distribution functions
Figure 4.3: Comparison of PMV side-chain contributions computed from the simulation to the experimental values obtained for the zwitterionic amino acids in aqueous buffer solution.

enables calculating the PMVs of a particular segment of an amino acid. Table 4.2 provides PMV contributions from the backbone and the side-chain atoms, including the $C_\alpha$ carbon atom. Note that these two contributions do not add up to the PMV of the amino acid as $C_\alpha$ atom was included in both the analysis. Also note that the backbone contributions are not identical, though the constituent groups of backbone are the same – it is highest for glycine and lowest for phenylalanine. This is so because, the solvent accessibility of the backbone atoms depends on the side-chain conformations and the bulkiness of the constituent atoms.

The PMV side-chain contribution can be determined as the difference between PMV of any amino acid and the PMV of glycine. Fig. 4.3 shows this difference obtained from the calculated value and compares them to the value obtained from the
Figure 4.4: The proximal effective grand canonical radial distribution functions for waters around proximal groups of the N-acetyl amino acid amides: methyl group (top panel), carbonyl oxygen group (bottom left panel) and nitrogen group (bottom right panel). Water proximal distributions around carbonyl oxygen in aspartic and glutamic acid and nitrogen group in lysine and arginine are distinct from others as shown in dashed lines.
Figure 4.5: Comparison of PMV side-chain contributions measured experimentally against values calculated by different approaches: 1D-RISM theory – circles, 3D-RISM theory – squares; simulations – triangles. The lines are linear fits through origin ($R^2 > 0.95$) with slopes of 0.54, 0.84 and 1.02 for 1D-RISM (dash-dotted line), 3D-RISM (dashed line) and simulations (solid line), respectively.

Experiments measuring the PMVs of amino acids in zwitter ionic form [119, 124, 125, 126]. Except for the charged amino acids– arginine, lysine, aspartic acid and glutamic acid – the agreement between calculated and experimental data is excellent. The observed discrepancies for the charged amino acids results from unscreened electrostatic interactions in simulations; counter ions were not added in the simulation.

The effect of charge can be clearly seen in the proximal radial distribution functions shown in Fig. 4.4. The proximal radial distribution of water around methyl groups in all 20 amino acids are identical. Whereas the water proximal distributions around nitrogen group in lysine and arginine and around carbonyl oxygen in aspartic and glutamic acid differ distinctly from other amino acids. Water is drawn closer
Figure 4.6: Comparison between the group contribution estimate of the PMV side-chain contributions and the experimental data for zwitterionic amino acid in aqueous buffer solution. Group contribution predictions are obtained using the group-averaged proximal correlation functions for uncharged amino acids (Eq. 4.8).
to the charged residues due to strong electrostatic interactions. Carbonyl oxygens of aspartic and glutamic acids are strongly hydrated than the amine groups of lysine and arginine [52].

Fig. 4.5 compares PMV side-chain contributions excluding lysine, arginine, aspartic and glutamic acid, calculated from the simulations to that estimated by 1D-RISM [132] and 3D-RISM [133] theories. The values predicted by 1D-RISM and 3D-RISM are on an average $\sim 54\%$ and $84\%$ of the experimental value. The calculations based on proximal radial distributions functions are within $2\%$ of the experimental data. Improved agreement with the inclusion of flexibility and side-chain conformations obtained in our approach, which was neglected by Hirata and coworkers [132, 133], further emphasis that hydration does not solely depend on the chemistry of the molecule, but also on the molecular topology.

The identical water proximal radial distribution around the constituent groups of the amino acids, excluding the four charged amino acids, suggests a PMV group contribution. The group averaged proximal correlation functions for group contribution method are determined by averaging proximal correlation functions of identical constituent groups across the 16 uncharged amino acids, as

$$
\langle g_{sw,\alpha}(r) \rangle = \frac{\sum_{\alpha}^{{\text{Identical groups}}} g_{sw,\alpha}(r) \frac{\partial V_{\alpha}(r)}{\partial r}}{\sum_{\alpha}^{{\text{Identical groups}}} \frac{\partial V_{\alpha}(r)}{\partial r}}.
$$

The PMVs for the uncharged amino acids computed from the group averaged proximal distributions are lower than that directly obtained from simulation (Table. 4.2) and the corresponding experimental values. Fig. 4.6 compares PMV side-chain contributions obtained from the group contribution method to experimental data. The
observed agreement is less accurate than that determined from individual simulation. The suggested group contribution technique, however, cannot explain the stability of different peptide conformations over others; e.g., the conformational preference of alanine-dipeptide for the polyproline II conformer over beta-2 conformer, which is due to the increased hydration of an amide group [145]. It has also been reported that the preferential hydration of amino acids on a protein surface varies between proteins and across the same protein surface depending on both the location of the amino acids and the environment [46]. The careful calculation of proximal volumes could accurately account for surface topology, but the cooperative hydration can be identified only case by case. Therefore, the group contribution technique provides only a qualitative description of amino acid or protein hydration.

4.5 Conclusions

The local density of water around an amino acid or a larger region can be determined from the proximal distribution of water around the constituent groups. We have demonstrated the utility of proximal distributions in estimating protein thermodynamic properties through KB theory by computing PMVs of the N-acetyl amino acid amides. Calculated PMVs are found to be consistently 5% higher than experimental values. The systematic error cancels out when PMV side-chain contributions are computed relative to glycine. PMV depends on the chemical nature of side-chain atoms and side-chain conformation, compromising the accuracy of applications of empirical group additivity for proteins purely on chemical considerations.
A group contribution scheme, if possible, should incorporate the local topology and chemical composition of the protein surface. Identical water proximal distributions around polar/non-polar molecular groups suggest a basis for group contribution scheme. However, extension to amino acids or protein hydration can be considered as only qualitative. The hydration of charged amino acids is distinctly different from polar/non-polar amino acids, and therefore, influence the hydration of adjacent polar/non-polar amino acids due to cooperative interactions.
CHAPTER 5

Preferential Interactions of Molecular Solutes and the Proximal Group Contributions

5.1 Introduction

Understanding the effects of co-solvents on the solvation of hydrophobic and hydrophilic solutes in aqueous solutions is fundamental to protein separations and purifications in the food [146, 147, 148, 149] and pharmaceutical [150, 151, 26, 152, 153, 154] industries. It is well known that co-solvents modify protein stability and solubility in aqueous solution [13, 155, 26]. However, the underlying molecular mechanisms of these effects are still not established [156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173]. Biomolecular processes, such as protein folding are dominantly driven by hydrophobic and hydrophilic interactions [62, 63, 64, 65], and co-solvent effects on these interactions can be dramatic.

Co-solvents typically used for protein separations and purifications span a wide range of chemical functionalities, including inorganic salts, organic solvents and long chain polymers [13, 155, 26]. In this work, we chose to study methanol, ethanol, glycerol and urea as co-solvents based on their biological significance. Alcohols can stabilize protein secondary structures, but destabilize protein native tertiary structures, leading to protein aggregation [174, 175]. Glycerol stabilizes protein native
tertiary structures, and is often used in protein crystallization to enhance the growth of protein crystals by reducing conformational flexibility \[155, 26\], while urea is the widely used protein-denaturant \[26, 166\].

### 5.2 Preferential Interaction Parameters

The co-solvent effects are described in terms of preferential interactions of co-solvent with the solute \[24, 26\]. These interactions are quantified by the preferential interaction parameter, $\Gamma_{sc}$. The macroscopic thermodynamic definition of $\Gamma_{sc}$ is the change in the chemical potential of the solute in response to a change in the co-solvent chemical potential:

$$\Gamma_{sc} \equiv -\left(\frac{\partial \mu_s}{\partial \mu_c}\right)_{T,P,m_s} = \left(\frac{\partial m_c}{\partial m_s}\right)_{T,P,\mu_c}$$ \hspace{1cm} (5.1)

where subscripts $c$ and $s$ refer to the co-solvent and solute, respectively, and $\mu_i$ and $m_i$ are the chemical potential and molal concentration of the component $i$. The second equality expresses this derivative in terms of measurable quantities \[162, 158, 159, 160\] as the change in the solute concentration for a given change co-solvent concentration.

Kirkwood-Buff (KB) solution theory \[31, 32\] provides a statistical thermodynamic framework for evaluating $\Gamma_{sc}$ from molecular distribution functions for the co-solvent and water (subscript $w$) around the solute:

$$\Gamma_{sc} = \rho_c (G_{sc} - G_{sw}) \hspace{1cm} (5.2)$$

where the KB integral, $G_{si}$, is derived from the radial distribution function for component $i$ around the solute, evaluated in the grand canonical ensemble:

$$G_{si} = 4\pi \int_0^\infty (g_{si}(r|\mu VT) - 1)r^2dr \hspace{1cm} (5.3)$$
In practice, it is convenient to evaluate the KB integrals from molecular simulations for a closed system at constant temperature and pressure or volume \[168, 169, 170, 171\] by defining a spherical sub-system, centered on the solute, within the larger closed system,

\[
G_{si}(R_c) \approx 4\pi \int_0^{R_c} (g_{si}(r) - 1)r^2 dr .
\] (5.4)

The cut-off radius, \(R_c\), is defined from the solute center, and in principle, corresponds to the distance beyond which \(g_{si}(r > R_c) \to 1\).

The approximate KB integral (Eq. 5.4) ignores the fact that the bulk density needed for normalizing the radial distribution function, \(g_{si}(r)\), will be different for open and closed simulation systems. An alternative definition, introduced by Matubayasi and Levy \[33\], corrects for the difference,

\[
G_{si}(R_c) = 4\pi \rho_i^{-1} \int_0^{R_c} \rho_{si}(r)r^2 dr - 4\pi \int_0^{R_c} r^2 dr .
\] (5.5)

Here, \(\rho_{si}(r)\) is the local density of the component \(i\) around the solute, and \(\rho_i\) is the bulk density of \(i\) and \(N_i\) is the total number of molecules of \(i\) in the system. By taking appropriate derivatives of Eqns. 5.3 and 5.5 an effective grand canonical radial distribution function, \(g^\Xi_{si}(R_c)\), is defined for the closed system \[134\], such that \(g^\Xi_{si}(R_c) \to 1\) when the density of the solvent \(i\) in the sub-system, \(\rho_{si}(r < R_c)\), equals the density in the rest of the system, \(\rho_{si}(r > R_c)\), or equivalently the corresponding excess chemical potential of the solvent are equal. \(\Gamma_{sc}\) (Eq. 5.2) is then defined in terms of the effective grand canonical radial distribution functions as,

\[
\Gamma_{sc}(R_c) = 4\pi \rho_c \int_0^{R_c} \left( g^\Xi_{sc}(r) - g^\Xi_{sw}(r) \right) r^2 dr .
\] (5.6)

The local preferential interaction parameter of the co-solvent with a constituent group, \(\alpha\), of the solute is evaluated from the proximal distributions of water and co-solvent
molecules around the constituent group,

\[
\Gamma_{sc,\alpha}(R_c) = \rho_c \int_0^{R_c} (g_{sc,\alpha}(r) - g_{sw,\alpha}(r)) \frac{\partial V_\alpha(r)}{\partial r} dr , \tag{5.7}
\]

where each proximal \( g_{si,\alpha}(r) \) is integrated over the corresponding proximal volume, \( V_\alpha \), for the constituent group \([137, 138]\). The sum of the local preferential interaction parameters, \( \Gamma_{sc,\alpha} \), over all constituent groups \( \alpha \) defining the solute reproduces the preferential interaction parameter for the entire solute,

\[
\Gamma_{sc}(R_c) = \sum_\alpha \Gamma_{sc,\alpha}(R_c) . \tag{5.8}
\]

Proximal distribution functions are determined by partitioning space around the solute into subdomains centered on each heavy atom \([134, 137, 138]\). For a molecular solute consisting of \( n \) sites, the proximal solute site \( \alpha \) at \( \vec{r}_\alpha \) for a solvent molecule at \( \vec{r} \) is determined by

\[
|\vec{r} - \vec{r}_\alpha| - \sigma_\alpha = \min_{\beta=1,\ldots,n} |\vec{r} - \vec{r}_\beta| - \sigma_\beta . \tag{5.9}
\]

Size disparities among solute sites are taken into account by defining an effective excluded-volume radius \( \sigma_\alpha \) for each solute site, \( \alpha = 1, \ldots, n \). This radius is computed as the distance at which the corresponding proximal radial distribution function first equals unity. For the two solutes considered here, neopentane and TMA, all four solvent accessible sites on each solute are the same size.

### 5.3 Smeared Distribution Functions

Long-range oscillations in the radial distribution functions are magnified in evaluating the KB integrals, which leads to poor convergence of \( \Gamma_{sc} \). Lockwood and Rossky \([34, 35]\) improved convergence of the KB integrals by “smearing” the distribution functions to dampen the long-range oscillations. Smearing takes advantage of
the fact that the KB integrals are independent of the choice of a reference point for
the solvent in defining its location with respect to the solute. The chosen reference
point may be one point, such as the center of mass, or as assumed here, uniformly
distributed within a spherical volume of radius, \(\lambda\), centered on the co-solvent molecule.

In addition, smearing the solvent distribution function within a single spherical
volume around the center of mass of the solvent is accurate for small molecules, such as
water and methanol, but is not likely to be accurate for larger co-solvents of practical
interest; e.g., ethanol, glycerol, or urea. For these co-solvents, smearing is carried
out within spherical volumes centered on the constituent groups of each co-solvent,
as shown in Fig. 5.1. Since the shape of the co-solvent molecules are considered using
this approach, orientational preferences of solvent molecules around the solute are
implicitly taken into account.

The resulting density distribution is the average over density distributions derived
from the reference points taken from this distribution:

\[
\rho(\vec{r}; \lambda) = \frac{\int_{|\vec{r}'| \leq \lambda} \rho(\vec{r} + \vec{r}') d\vec{r}'}{4\pi \lambda^3 / 3}.
\]

(5.10)

The vector \(\vec{r}'\) is defined relative to the solvent center at \(\vec{r}\), and \(\rho(\vec{r} + \vec{r}')\) is the local
density at a distance \(\vec{r} + \vec{r}'\) from the solute defined using \(\vec{r}'\) as the reference point for
each solvent molecule. The radial distribution of local densities are obtained by angle
averaging over all solvent positions.

For co-solvent molecules consisting of multiple heavy atoms or atom groups, uniform
distributions of reference points are accordingly defined for each constituent
group \(j\). For a co-solvent molecule of \(m\) constituent groups,

\[
\rho(\vec{r}; \lambda_1, \ldots, \lambda_m) = \frac{1}{m} \sum_{j=1}^{m} \rho_j(\vec{r}; \lambda_j).
\]

(5.11)
Figure 5.1: Concept of smearing the co-solvent distribution function: The location of a co-solvent molecule – in this case, urea – relative to the solute of interest is defined by a randomly selected reference point within the shaded spherical volumes centered on the different constituent groups of urea, rather than a single reference point; e.g., the center of mass. Radii defining the spherical volumes for the different constituent groups are: 2.3 Å for an oxygen atom or hydroxyl group, 2.5 Å for a nitrogen atom or amine group, and 3.0 Å for a carbon atom or a methyl or methylene group. These radii are the radial distances up to which the corresponding proximal distribution functions equal zero.

5.4 Simulation Details

The united atom approach was used to model all solutes: methane, ethane, neopentane and TMA, and co-solvents such as methanol, ethanol and urea. The atomic charges and Lennard-Jones (LJ) parameter values were taken from the OPLS model [176, 177, 178, 179]. The intramolecular interactions were included as the harmonic potentials for bond lengths and bond angles and a fourier series potential of the form \( \Sigma (V_n/2)(1 + \cos(n\phi)) \) for the torsional motions [176, 178, 180]. For glycerol, intramolecular potential for the united atom model was not completely available. Glycerol was, therefore, modeled by all atom approach with the CHARMM35 forcefield [181]. TIP3P model was used for water [140].
Table 5.1: Number of co-solvent and water molecules in the simulation as a function of the co-solvent concentration

<table>
<thead>
<tr>
<th>co-solvent conc. mol fraction</th>
<th>No. of co-solvent molecules</th>
<th>No. of water molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0</td>
<td>4096</td>
</tr>
<tr>
<td>0.05</td>
<td>205</td>
<td>3891</td>
</tr>
<tr>
<td>0.10</td>
<td>410</td>
<td>3686</td>
</tr>
<tr>
<td>0.20</td>
<td>820</td>
<td>3276</td>
</tr>
<tr>
<td>0.30</td>
<td>1229</td>
<td>2867</td>
</tr>
<tr>
<td>0.50</td>
<td>2048</td>
<td>2048</td>
</tr>
<tr>
<td>1.00</td>
<td>4096</td>
<td>0</td>
</tr>
</tbody>
</table>

Each simulation box consisted of a solute solvated in a well equilibrated aqueous solution of co-solvent. The number of molecules of co-solvent and water for a given co-solvent concentration is listed in Table 5.1. All simulations were carried out using NAMD 2.6 [58] with an integration time step of 2 fs. The system was initially minimized for 10,000 steps and then equilibrated for 3 ns at isothermal-isobaric condition of 300 K and 1 bar pressure. The average volume of the system determined from the last 2 ns of NPT simulation was used for the subsequent Canonical simulation. System was further equilibrated for 1 ns at NVT condition at 300 K. Configurations were saved every 0.05 ps over the 5 ns of NVT production run.

Temperature was held constant in the simulations by applying Langevin dynamics to all heavy atoms using a damping coefficient of 1 ps$^{-1}$. Constant pressure was maintained using a Nosé-Hoover Langevin piston [141, 142] with a period of 200 fs and a decay of 100 fs. Periodic boundary conditions were imposed, and the particle-mesh Ewald method [143] with a real-space cutoff of 12 Å was used in computing the electrostatic interactions. The same cutoff was applied to nonbonded nonelectrostatic
interactions. The positions of all hydrogen atoms in the system were constrained using the SHAKE algorithm [60].

5.5 Results and Discussion

The effect of smearing and the closed system correction for the methanol (co-solvent) and water distribution functions is shown in Fig. 5.2 where $\Gamma_{sc}(R_c)$ for methanol with neopentane (solute) in the 20 mol% aqueous methanol solution is calculated using smeared effective grand canonical radial distributions and unsmeared canonical radial distributions. Using canonical radial distributions, $\Gamma_{sc}(R_c)$ decreases initially, and then abruptly increases to a maximum of $\sim 2.5$ between 6 - 7 Å, before
steadily decreasing without converging for $R_c > 10$ Å. The negative $\Gamma_{sc}$ for small $R_c$ indicates preferential exclusion of methanol near neopentane, and simply reflects the larger excluded volume of methanol relative to water. The lack of convergence, indeed the divergence at larger $R_c$ is attributed to underdamped, long-range oscillations in the water and methanol density distribution functions that are also out of phase because of the different characteristic length scales of water and methanol density fluctuations. Similar behavior has been reported in other simulation studies of preferential interactions for molecular solutes [168, 169, 170, 171].

When smeared solvent distribution functions are used, $\Gamma_{sc}(R_c)$ increases monotonically starting at about 2 Å up to a maximum at $\sim 8$ Å, and then converges to a value slightly above +2.0 for $R_c \geq 12$ Å. The effect of different excluded volumes for methanol and water that leads to $\Gamma_{sc}(R_c) < 0$ at small $R_c$ without smearing is eliminated by using radii of 2.3 Å and 3.3 Å to smear the water and methanol distribution functions, respectively. $R_c = 2.3$ Å corresponds to the effective hard-sphere diameter for TIP3P water oxygens, while $R_c = 3.3$ Å defined with respect to the methanol center of mass was chosen to give an accurate estimate of the partial molar volume at infinite dilution for methane in methanol [182] using KB theory.

Fig. 5.3 shows $\Gamma_{sc}(R_c)$ for methanol with neopentane and TMA at different aqueous methanol concentrations calculated using the smeared distribution functions for methanol and water. Reasonable convergence is achieved for $R_c > 12$ Å over the entire concentration range for both solutes, such that $\Gamma_{sc}$ can be reliably calculated as an average value over $R_c = 12 - 15$ Å. The results are reported in Table 5.2. For neopentane, $\Gamma_{sc}$ is greater than zero and increases by about an order of magnitude.
Figure 5.3: Preferential interaction parameters for methanol (co-solvent) with the solutes, neopentane (top panel) and TMA (bottom panel), in aqueous solutions at different methanol concentrations as a function of cutoff radius, $R_c$. $\Gamma_{sc}$ is calculated using the smeared effective grand canonical distributions for methanol and water. Methanol concentrations (mol fraction): 0.05 - solid line; 0.10 - dotted line; 0.20 - dashed line; 0.30 - dash-dot line; 0.50 - dash-dot-dash line.
Table 5.2: Preferential interaction parameters for methanol (co-solvent) with solutes, neopentane and TMA, in aqueous methanol solutions, evaluated as the sum of preferential interaction parameters for the constituent groups of the solutes. Standard deviations obtained by dividing the simulation trajectory into five equal parts and block averaging.

<table>
<thead>
<tr>
<th>Methanol concentration (mol fraction)</th>
<th>Neopentane</th>
<th>TMA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Gamma_{sc}$</td>
<td>$\sum_{\alpha} \Gamma_{sc,\alpha}$</td>
</tr>
<tr>
<td>0.05</td>
<td>0.45 ± 0.17</td>
<td>0.44 ± 0.16</td>
</tr>
<tr>
<td>0.10</td>
<td>1.22 ± 0.27</td>
<td>1.16 ± 0.31</td>
</tr>
<tr>
<td>0.20</td>
<td>2.06 ± 0.21</td>
<td>2.00 ± 0.28</td>
</tr>
<tr>
<td>0.30</td>
<td>3.26 ± 0.29</td>
<td>3.25 ± 0.40</td>
</tr>
<tr>
<td>0.50</td>
<td>4.58 ± 0.42</td>
<td>4.50 ± 0.37</td>
</tr>
</tbody>
</table>
with increasing methanol concentration up to 50 mol% methanol, the highest concentration considered, indicating the preferential accumulation of methanol relative to water near the solute. In contrast, $\Gamma_{sc}$ for methanol with TMA is less than zero and becomes more negative by about an order of magnitude with increasing methanol concentration over the same concentration range, indicating the preferential accumulation of water relative to methanol near this solute.

Also reported in Table 5.2 are methanol-neopentane and methanol-TMA preferential interaction parameters calculated as the sum of preferential interaction parameters for methanol with the constituent methyl groups of neopentane and TMA (Eq. 5.8) using smeared proximal water and methanol distribution functions around each group. $\Gamma_{sc}$ obtained by group additivity matches within statistical uncertainty the $\Gamma_{sc}$ calculated directly for the entire solute at each methanol concentration. The accuracy of group additivity for neopentane is not surprising, since the methyl groups are not charged [136]. Evidently, the symmetry of the charged methyl groups for TMA results in the accuracy of group additivity for this solute. Although group additivity is not expected to hold for charged solutes or co-solvents when the charges are unevenly distributed, group additivity of solute sites coupled with smeared proximal water and co-solvent distribution functions can be implemented with some care to compute preferential interaction parameters for much larger solutes than those considered here.

Co-solvent preferential interaction parameters as a function of $R_c$ for ethanol, glycerol, and urea with neopentane and TMA in 20 mol% aqueous co-solvent solutions, computed by smearing the solvent distribution functions based on their constituent groups, are shown in Fig. 5.4. $\Gamma_{sc}$ calculated as the average value of $\Gamma_{sc}(R_c)$ for
Figure 5.4: Preferential interaction parameters for different co-solvents with the solutes, neopentane (top panel) and TMA (bottom panel), in 20 mol % aqueous co-solvent solutions as a function cutoff radius, $R_c$, calculated using smeared effective grand canonical distribution functions for the co-solvents and water. Co-solvents: methanol - solid line; ethanol - dotted line; glycerol - dash-dot line; urea - dash-dot-dash line. The horizontal dashed line is the baseline corresponding to $\Gamma_{sc} = 0$. The co-solvent distributions were smeared based on their constituent groups.
Table 5.3: Preferential interaction parameters for different co-solvents with solutes, neopentane and TMA in 20 mol % aqueous co-solvent solutions, evaluated by smearing the co-solvent effective grand canonical distribution around the constituent groups of each co-solvent.

<table>
<thead>
<tr>
<th>Co-solvent</th>
<th>$\Gamma_{sc}$ Neopentane</th>
<th>$\Gamma_{sc}$ TMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethanol</td>
<td>8.44 ± 1.02</td>
<td>-3.02 ± 0.80</td>
</tr>
<tr>
<td>Glycerol</td>
<td>-0.07 ± 0.69</td>
<td>-3.38 ± 0.66</td>
</tr>
<tr>
<td>Urea</td>
<td>0.67 ± 0.54</td>
<td>-1.32 ± 0.54</td>
</tr>
</tbody>
</table>

$R_c > 12 \, \text{Å}$ is given in Table 5.3. In comparing ethanol to methanol (Table 5.2) as co-solvents for neopentane, we find that $\Gamma_{sc}$ for ethanol is about four times greater than that for methanol at this co-solvent concentration. The much larger $\Gamma_{sc}$ for ethanol is attributed to more favorable hydrophobic interactions owing to the additional methylene group of ethanol. In contrast, there is no significant difference between ethanol and methanol as co-solvents for TMA, which is preferentially hydrated in either case. Glycerol shows no preferential interactions with neopentane with $\Gamma_{sc} \sim 0$, and is preferentially excluded by TMA to the same extent as methanol and ethanol. Urea exhibits much weaker preferential interactions as a co-solvent for neopentane and TMA, although $\Gamma_{sc}$ is positive for neopentane and negative for TMA. This behavior is attributed to the high dipole moment ($\sim 4$ Debye) and small molecular volume of urea compared to ethanol and glycerol.
5.6 Conclusions

Preferential interaction parameters for methanol, ethanol, glycerol, and urea (co-solvents) with neopentane and tetra-methyl ammonium ion (solutes) in aqueous solutions were evaluated from molecular simulations using KB solution theory. Improved convergence of the KB integrals are obtained by smearing the molecular distribution functions for water and the co-solvent around the solute. For the larger co-solvents – ethanol, glycerol, and urea – smearing is applied on the basis of the constituent groups comprising these molecules to enable better sampling of all possible co-solvent orientations around the solute. We find that neopentane is preferentially solvated by methanol, and to a much greater extent by ethanol, but this hydrophobic solute shows no preference for glycerol or urea over water. In constrast, TMA is preferentially hydrated is the presence of all the co-solvents, although to a lesser extent for urea compared to the other three co-solvents. Interestingly, the preferential interaction parameters for methanol, ethanol, and glycerol with TMA are essentially the same, indicating the solute is preferentially hydrated to the same extent in the presence of all three co-solvents.

Preferential interaction parameters for neopentane and TMA were also computed as the sum of four, equivalent solute site-specific preferential interaction parameters, derived from proximal water and co-solvent density distributions around the constituent methyl group for each solute. These preferential interaction parameters were found to be accurately group additive, suggesting that co-solvent preferential interactions can be estimated for much larger solutes in the absence of significant cooperative solvation effects among neighboring groups. Recognizing that group additivity is not expected to hold in general for complex, chemically heterogeneous solutes, such as
globular proteins, these results offer the alternative possibility of evaluating preferential interaction parameters locally across the surface of a protein, in essence to map co-solvent preferential interactions as a function of specific sites on the protein surface.
CHAPTER 6

Quasi-Chemical Theory of Preferential Interactions in Aqueous Solutions

6.1 Introduction

The solvation of apolar solutes in aqueous solutions, and specifically the effects of co-solvents on solvation, are fundamentally important to understanding the thermodynamic driving forces that influence biomolecular processes, such as self-assembly and molecular recognition. Co-solvent effects are described in terms of their preferential interactions with the solute of interest. The macroscopic thermodynamic definition of these preferential interactions is given by the change in the chemical potential of the solute, $\mu_s$, in response to a change in co-solvent chemical potential, $\mu_c$, at constant temperature, pressure, and solute concentration \[161, 183, 25\]. The preferential interaction parameter,

$$\Gamma_{sc} \equiv - \left( \frac{\partial \mu_s}{\partial \mu_c} \right)_{T, P, \rho_s} = \left( \frac{\partial \rho_c}{\partial \rho_s} \right)_{T, P, \mu_c},$$

(6.1)

quantifies these interactions in terms of experimentally accessible changes in co-solvent and solute concentrations \[162, 158, 159, 160, 157, 184\], here the solute and co-solvent number densities, $\rho_s$ and $\rho_c$, in the second equality.
Co-solvent effects are characterized at the molecular level in terms of preferential intermolecular interactions with the solute. A rigorous statistical mechanical expression for the preferential interaction parameter in the limit of an infinitely dilute solute is derived from Kirkwood-Buff (KB) solution theory [185],

\[
\Gamma_{sc} = \rho_c(G_{sc} - G_{sw}),
\]

(6.2)

where the KB integrals, \(G_{sc}\) and \(G_{sw}\), are obtained from the co-solvent-solute and water-solute radial distribution functions (rdfs) evaluated for an open system,

\[
G_{si} = 4\pi \int_0^\infty (g_{si}(r|\mu VT) - 1) r^2 dr.
\]

(6.3)

Eq. 6.2 provides a molecular view of preferential interactions in terms of the local concentrations of co-solvent and water molecules around the solute relative to their concentrations in bulk solution. For preferential partitioning of co-solvent molecules over water molecules in the vicinity of the solute, \(\Gamma_{sc} > 0\), while preferential partitioning of water molecules over co-solvent molecules in the vicinity of the solute corresponds to \(\Gamma_{sc} < 0\).

Simulation studies of molecular preferential interactions typically approximate the KB integrals in Eq. 6.3 by defining an open subsystem within a closed simulation system at constant temperature and pressure or volume [170, 169, 163, 186]. For a spherical subsystem defined by the radius, \(R\), from the solute center,

\[
G_{si} \approx 4\pi \int_0^R (g_{si}(r) - 1) r^2 dr
\]

(6.4)

will be reasonably accurate if \(g_{si}(r)\), the constant NPT or NVT rdf, is evaluated for \(R\) up to two or three solvation shells around the solute. The preferential interaction parameter, however, is obtained from the difference in KB integrals for the co-solvent,
and water, and can be sensitive the upper limit of integration in this approximation \[187, 186\]. For large solutes, such as globular proteins, $\Gamma_{sc}$ averaged over the entire solute-solution interface is found to converge at distances of 6-8 Å into the solution \[188, 166, 189, 184\]. In contrast, $\Gamma_{sc}$ converges only at very large $R$ for small molecular solutes, and therefore, inordinately large simulation systems are required for reliable estimates of $\Gamma_{sc}$ for these solutes \[186, 190, 191, 163, 192\].

New force fields have been proposed to improve convergence of the KB integrals for specific co-solvents and molecular solutes in aqueous solution \[190, 191, 163, 192\]. An alternative approach improves convergence by “smearing” the positions of the solvent molecules, thereby suppressing long-range correlations in the solute-solvent rdfs \[34\]. This approach has been successfully applied to computing partial molar volumes at infinite dilution for a variety of small-molecule solutes \[34, 35, 134\], and for molecular solutes consisting of multiple constituent groups based on proximal rdfs around the individual constituent groups \[35, 134\].

In this work, we describe an entirely different approach for evaluating $\Gamma_{sc}$ from molecular simulations based on quasi-chemical (QC) theory and the potential distribution theorem (PDT). QC theory \[29, 193, 194\] distinguishes solute-solvent interactions within a proximal region or inner shell around the solute from solute interactions with the solvent outside this inner shell. Inner-shell interactions are described by the chemical association of a solute molecule with molecular clusters of the solvent averaged over a statistical distribution of solvent compositions. Solute interactions with outer-shell solvent molecules make an important contribution by preferentially stabilizing certain inner-shell solvent clusters over others.
We apply QC theory and the PDT to compute $\Gamma_{sc}$ for the simplest hydrophobic solute, methane, in aqueous methanol solutions, and show that this preferential interaction parameter is independent of the definition of the inner shell. Our results emphasize the importance of solute interactions with outer-shell solvent, and more significantly, permits defining an inner shell that is much smaller in size than proximal region required to evaluate $\Gamma_{sc}$ from KB integrals when the outer-shell contribution is taken into account. Since less stringent demands are placed on the molecular details of these solute-solvent interactions, the outer-shell contribution to $\Gamma_{sc}$ can be accurately estimated using approximate solution thermodynamic theories, which is an advantage of the QC/PDT methodology for molecular solutes. A second, potential advantage is to enable $\Gamma_{sc}$ to be computed locally by defining inner shells centered on specific sites on much larger macromolecular solutes, such as proteins and protein assemblies, to give two notable examples.

6.2 Quasi-Chemical Theory

QC theory and the PDT have been successfully applied to computing hydration free energies from molecular simulations for a variety of solute, including ions [195, 196, 197, 198, 199, 200], polar solutes, including water [112, 113, 136], and hydrophobic solutes [201, 202, 203]. We briefly review QC theory before applying it to co-solvent preferential interactions with the solute in aqueous solutions.

The solute chemical potential or the partial molar Gibbs free energy of the solute, $\mu_s$, in solution is defined by

$$\mu_s = kT \ln \rho_s \Lambda_s^3 + \mu_{s}^{ex}.$$  \hfill (6.5)
The first term on the right side of this expression is the ideal contribution, consisting of the thermal energy, $kT$ (product of Boltzmann’s constant $k$ and temperature), solute concentration, $\rho_s$, defined here as the number density of solute molecules in solution, and the solute thermal deBroglie wavelength, $\Lambda_s$. The chemical potential in excess of this ideal contribution, $\mu_s^{ex}$, accounts for intermolecular interactions among the three species in solution.

The solute excess chemical potential is defined here by the inverse form of the PDT [29],

$$\mu_s^{ex} = kT \ln \langle e^{+\beta \epsilon} \rangle,$$

where $\epsilon$ is the interaction or binding energy of the solute to the solvent, and the brackets indicate thermal averaging over all possible solute-solvent configurations in the fully coupled system at temperature, $T = 1/k\beta$.

Intermolecular solute-solvent interactions are partitioned into local interactions within a defined proximal volume or inner shell around the solute, and solute interactions with solvent molecules outside this inner shell – i.e., the outer shell corresponding to the rest of the system. Inner shell occupancy is specified by defining an indicator function $b_j$, such that $b_j = 1$ when solvent molecule $j$ occupies the inner shell, and $b_j = 0$ when the solvent molecule is outside that volume. With this definition, the solute excess chemical potential is defined by

$$\beta \mu_s^{ex} = \ln x_0 + \ln \left( e^{\beta \epsilon} \prod_j (1 - b_j) \right),$$

with $x_0$ the probability of finding configurations with no solvent in the inner shell.

The occupancy probability $x_0$ is obtained by defining the coordination state of the solute in terms of the stoichiometry for chemical association of the solute ($S$) with $n$
water ($W$) and $m$ co-solvent molecules ($C$) in the inner shell,

$$S + nW + mC \rightleftharpoons SW_n C_m.$$  

The occupancy probability $x_0$ is thus given by

$$x_0 = \frac{c_{00}}{\sum_n \sum_m c_{nm}} = \frac{1}{1 + \sum_{n \geq 1} \sum_{m \geq 1} K_{nm} \rho_w^n \rho_c^m},$$

where the concentration $c_{nm}$ corresponds to a cluster of $n$ water and $m$ co-solvent molecules with the solute, $K_{nm}$ is the “products over reactants” equilibrium constant for chemical association (Eq. 6.8), and $\rho_w$ and $\rho_c$ are water and co-solvent densities in bulk solution, respectively.

The outer-shell interaction contribution on the right side of Eq. 6.7 can be separated into one contribution that accounts for intermolecular interactions between the solute and solvent in the outer shell under the condition that no solvent molecules occupy the inner shell, and another to account for the free energy of creating a cavity in the solvent corresponding to the inner shell,

$$\ln \left\langle e^{+\beta \epsilon} \prod_j (1 - b_j) \right\rangle = \ln \left\langle e^{\beta \epsilon} | n, m = 0 \right\rangle - \ln p_0.$$  

The free energy of cavity formation, $-kT \ln p_0$, is expressed in terms of the probability, $p_0$, of finding a cavity in solvent the size and shape of the inner shell. This probability is likewise defined by the chemical association in Eq. 6.8 but without the solute present:

$$p_0 = \frac{1}{1 + \sum_{n \geq 1} \sum_{m \geq 1} \tilde{K}_{nm} \rho_w^n \rho_c^m}.$$
The tilde over the equilibrium constant, $\tilde{K}_{nm}$, denotes that the cluster of $n$ water and $m$ co-solvent molecules is formed in the absence of solute in the inner shell. Thus, the solute excess chemical potential (Eq. 6.7) becomes

$$\beta \mu^e_s = \ln x_0 - \ln p_0 + \ln \langle e^{\beta \epsilon |n, m = 0} \rangle.$$  \hspace{1cm} (6.12)$$

If the conditional probability distribution of solute binding energies is accurately gaussian [113, 202, 203, 136], the outer-shell contribution is defined by the mean and the variance of this distribution, and

$$\beta \mu^e_s = \ln x_0 - \ln p_0 + \beta \langle \epsilon \rangle + \beta^2 \sigma^2/2.$$  \hspace{1cm} (6.13)$$

The preferential interaction parameter expressed as a partial derivative of $\beta \mu^e_s$ is derived from Eq. 6.1 assuming $\partial \beta \mu_c / \partial \rho_c \propto \rho_c^{-1}$,

$$\Gamma_{sc} = -\rho_c \left( \frac{\partial \beta \mu^e_s}{\partial \rho_c} \right)_{T,P,\rho_s}.$$  \hspace{1cm} (6.14)$$

Evaluating this partial derivative using the expression for the solute excess chemical potential given by Eq. 6.12 gives

$$\Gamma_{sc} = -\rho_c \left( \frac{\partial \ln x_0}{\partial \rho_c} \right)_{T,P,\rho_s} + \rho_c \left( \frac{\partial \ln p_0}{\partial \rho_c} \right)_{T,P,\rho_s}$$

$$- \rho_c \left( \frac{\partial \ln \langle e^{\beta \epsilon |n, m = 0} \rangle}{\partial \rho_c} \right)_{T,P,\rho_s}.$$  \hspace{1cm} (6.15)$$

The first two terms on the right side of this equation describe inner-shell preferential interactions in terms of the co-solvent effect on the composition of the solvent molecular clusters that associate with the solute in the inner shell. Both chemical equilibrium for solute-solvent association (Eq. 6.8) and solvent molecular packing around the solute contribute to these preferential interactions. The rightmost term describes the co-solvent effect on outer-shell solvent interactions with the solute. This contribution...
recognizes that the formation of inner-shell solvent molecular clusters with the solute is stabilized by the outer-shell solvent, and thus, certain solvent molecular clusters are preferentially stabilized over others depending on the bulk concentration of co-solvent.

Inner-shell preferential interactions are evaluated by simply accounting for differences in the average number of water molecules, \( \langle n \rangle \), and co-solvent molecules, \( \langle m \rangle \), occupying the inner shell with the solute, relative to these average occupancy numbers in the absence of solute, \( \langle n \rangle_0 \) and \( \langle m \rangle_0 \), and scaling by the ratio of co-solvent and water partial molar volumes, \( \bar{V}_c/\bar{V}_w \), to account for their different molecular volumes in solution:

\[
\Gamma_{sc}(\text{inner shell}) = \langle m \rangle - \langle m \rangle_0 - \left( \langle n \rangle - \langle n \rangle_0 \right) \frac{\rho_c \bar{V}_c}{\rho_w \bar{V}_w}.
\]  

Eq. 6.16 assumes the chemical equilibrium constants, \( K_{nm} \) and \( \tilde{K}_{nm} \), do not depend on the co-solvent concentration – i.e., the co-solvent and water form an ideal mixture. If the difference in co-solvent and water partial molar volumes is neglected,

\[
\Gamma_{sc}(\text{inner shell}) = \langle m \rangle - \langle n \rangle \frac{\langle m \rangle_0}{\langle n \rangle_0},
\]  

which is similar to the expressions for the preferential interaction parameter derived from KB theory \[188\] \[184\], except that all quantities in Eq. 6.17 are defined exclusively within the inner shell around the solute. We note that this preferential interaction parameter depends on the size and shape of the inner shell. However, the full QC/PDT expression for \( \Gamma_{sc} \) (Eq. 6.15) is independent of the inner-shell definition, which suggests that outer-shell contribution, the rightmost term in Eq. 6.15, is important for an accurate determination of the preferential interaction parameter,
and as shown below, allows much smaller inner shell volumes to be defined compared to those required to obtain $\Gamma_{sc}$ from KB integrals (Eq. 6.2).

## 6.3 Simulation Details

All simulations were carried out using NAMD 2.6 [58] with an integration time step of 2 fs. The numbers of water and methanol molecules in each simulation system are given in Table 6.1. United atom OPLS models [176, 178] were used for methane and methanol, and the TIP3P model was used for water [140].

Each system was initially relaxed by executing 10,000 steps of conjugate gradient energy minimization, and then equilibrated for 3 ns at 300 K and 1 bar. The simulation box size was then fixed to the average box size for the final 2 ns of this constant NPT simulation, and the system further equilibrated for 1 ns at constant volume and 300 K. Configurations were saved every 0.1 ps over a 25-ns constant NVT production run.

During the simulation, the system temperature was held constant by applying Langevin dynamics to all heavy atoms using a damping coefficient of 1 ps$^{-1}$. Constant pressure was maintained using a Nosé-Hoover Langevin piston [141, 142] with a period of 200 fs and a decay of 100 fs. Periodic boundary conditions were imposed, and the particle-mesh Ewald method [143] with a real-space cutoff of 12 Å was applied in computing the electrostatic interactions. The non-bonded dispersion interactions were smoothly switched to zero at 12 Å starting from 10 Å. The SHAKE algorithm [60] was used to constrain the positions of all hydrogen atoms.
Table 6.1: System Specifications

<table>
<thead>
<tr>
<th>System</th>
<th>number of methane molecules</th>
<th>number of methanol molecules</th>
<th>number of water molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td>water</td>
<td>0/1</td>
<td>0</td>
<td>1728/4096*</td>
</tr>
<tr>
<td>methanol</td>
<td>0 or 1</td>
<td>4096</td>
<td>0</td>
</tr>
<tr>
<td>5 mol% methanol/water solution</td>
<td>0 or 1</td>
<td>205</td>
<td>3891</td>
</tr>
<tr>
<td>10 mol% methanol/water solution</td>
<td>0 or 1</td>
<td>410</td>
<td>3686</td>
</tr>
<tr>
<td>20 mol% methanol/water solution</td>
<td>0 or 1</td>
<td>820</td>
<td>3276</td>
</tr>
<tr>
<td>30 mol% methanol/water solution</td>
<td>0 or 1</td>
<td>1229</td>
<td>2867</td>
</tr>
</tbody>
</table>

*second number corresponds to methane in water.
Figure 6.1: Radial distribution functions for water oxygens around a methane molecule in water (solid line) and in aqueous methanol solutions at different methanol mole fractions: 0.05 (dashed line), 0.10 (dash-dot line), 0.20 (dotted line), 0.30 (long dashed line). Radial distribution functions are shifted successively upwards by 0.2 units for clarity.
Figure 6.2: Radial distribution functions for methanol methyl groups (top) and hydroxyl groups (bottom) around a methane molecule in methanol (solid line) and in aqueous methanol solutions at different methanol mole fractions: 0.05 (dashed line), 0.10 (dash-dot line), 0.20 (dotted line), 0.30 (long dashed line). Radial distribution functions are shifted successively upwards by 0.2 units for clarity.
6.4 Results and Discussion

Fig. 6.1 shows rdfs for water oxygens around a methane molecule in water and in aqueous methanol solutions over the range of methanol concentrations considered. Adding methanol to the solution decreases the height of the primary peak, which nonetheless remains above 1.0 over the entire range of methanol concentrations. In addition, the peak position is unchanged over this concentration range. The results allow an inner shell for water molecules around methane to be defined that is accurately independent of methanol concentration. This inner shell is defined as a sphere centered on methane with a radius corresponding to the distance at which the water oxygen-methane rdf first equals 1.0. At this distance, methane-water repulsive interactions are counterbalanced by methane-water attractive interactions. From Fig. 6.1, the inner-shell radius for water molecules around a central methane molecule is 3.3 Å, independent of methanol concentration.

Fig. 6.2 shows rdfs for both the methyl and hydroxyl groups of methanol around a methane molecule in methanol and in aqueous methanol solutions over the same range of methanol concentrations. The primary peaks for the methyl group rdfs are higher, sharper, and closer to methane compared to the corresponding primary peaks for the hydroxyl group rdfs, indicating that near-neighbor methanol molecules are oriented, such that their methyl groups point towards methane. The primary peak heights in each case remain above 1.0 over the entire range of methanol concentrations, and the peak positions are independent of methanol concentration.

Two radii are required to specify a spherical inner shell for methanol molecules around a central methane molecule, one for the methyl group and another for the hydroxy group [80]. As above, the radii are defined as the distance at which the
Table 6.2: QC contributions to $\mu_s^{ex}$ (Eq. 6.13) for methane in aqueous methanol solutions as a function of methanol mole fraction computed using mean binding energies and fluctuation contributions obtained from gaussian descriptions of the conditional probability distributions in Figs. 6.3 and B.3 with inner-shell radii: $\lambda_w = 3.3$, $\lambda_m = 3.6$, and $\lambda_h = 3.9$ Å. All free energies are in kcal/mol.

<table>
<thead>
<tr>
<th>methanol mole frac.</th>
<th>$-kT \ln p_0$</th>
<th>$kT \ln x_0$</th>
<th>$\langle \varepsilon \rangle$</th>
<th>$\beta \sigma^2 / 2$</th>
<th>$\mu_s^{ex}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>+6.34</td>
<td>-0.36</td>
<td>-4.06</td>
<td>+0.12</td>
<td>+2.04</td>
</tr>
<tr>
<td>0.05</td>
<td>+6.20</td>
<td>-0.45</td>
<td>-4.04</td>
<td>+0.13</td>
<td>+1.84</td>
</tr>
<tr>
<td>0.10</td>
<td>+6.15</td>
<td>-0.52</td>
<td>-4.01</td>
<td>+0.14</td>
<td>+1.76</td>
</tr>
<tr>
<td>0.20</td>
<td>+5.95</td>
<td>-0.62</td>
<td>-3.94</td>
<td>+0.15</td>
<td>+1.54</td>
</tr>
<tr>
<td>0.30</td>
<td>+5.77</td>
<td>-0.70</td>
<td>-3.90</td>
<td>+0.16</td>
<td>+1.33</td>
</tr>
<tr>
<td>1.00</td>
<td>+4.63</td>
<td>-0.78</td>
<td>-3.60</td>
<td>+0.17</td>
<td>+0.42</td>
</tr>
</tbody>
</table>

corresponding rdfs in Fig. 6.2 first equal 1.0. These radii are 3.6 Å and 3.9 Å for the methanol methyl and hydroxyl groups, respectively, independent of methanol concentration.

Fig. 6.3 shows the conditional probability distributions of binding energies for methane in water and in methanol when no solvent occupies the inner shell around methane. Conditioning gives a nearly gaussian distribution in both cases by eliminating close contact, repulsive interactions between methane and solvent molecules that contribute to the high energy tail of the probability distribution. Larger deviations from gaussian behavior are observed for methane in water at high binding energies. However, the estimated error in $\mu_s^{ex}$ introduced by the gaussian approximation is on the order of $10^{-2}$ kcal/mol [136], which is comparable to the statistical uncertainty in sampling the binding energies. See Table B.1 and associated text in Appendix B.
Figure 6.3: Conditional probability distributions of binding energies for methane in water (solid circles) and in methanol (open circles) when no water oxygens are within $\lambda_w = 3.3$ Å of methane in water, and no methanol methyl and hydroxyl groups are within $\lambda_m = 3.6$ Å and $\lambda_h = 3.9$ Å, respectively, of methane in methanol. Lines are gaussian fits to the data.
The QC contributions to $\mu_{s}^{ex}$ for methane in water, methanol, and the aqueous methanol solutions considered are collected in Table 6.2. The dominant contributions in all cases come from the unfavorable free energy of cavity formation, which is compensated in large part by the favorable mean binding energy of methane to outer-shell solvent. The free energy contributions arising from chemical association and fluctuations in the binding energy are an order of magnitude smaller, and also partially offset one another.

The calculated free energy of methane hydration is 2.04 kcal/mol, compared to $\mu_{s}^{ex} = 2.23$ kcal/mol obtained by thermodynamic integration in a previous simulation study of united atom, OPLS methane in TIP3P water [204]. The calculated value is also in agreement with measured values that range from 1.94 to 2.01 kcal/mol [205, 11, 206]. The calculated free energy of methane solvation in methanol is 0.42 kcal/mol, which is virtually equal to the measured value of 0.39 kcal/mol [205]. From Table 6.2 the higher methane solubility in methanol is attributed almost entirely to the less unfavorable free energy of cavity formation in the organic solvent, as expected [207, 208], even though the defined inner shell is larger.

QC contributions to $\mu_{s}^{ex}$ for methane in water and in methanol as a function of the defined inner-shell radii for the two solvents are plotted in Fig. 6.4. In both cases, $\mu_{s}^{ex}$ is independent of the definition of the inner shell. In contrast, the free energies of cavity formation and chemical association are sensitive to the defined inner shell – i.e., more work is required to create larger cavities in either solvent, but the solute-solvent molecular clusters become more stable as the inner shell expands to included more solvent. However, the two free energy contributions compensate one another to such an extent that their net contribution to $\mu_{s}^{ex}$ does not change significantly as the inner
Figure 6.4: Top Panel: QC contributions to $\mu_s^{ex}$ (Eq. 6.13) for methane in water as a function of the inner-shell radius, $\lambda_w$, for the water oxygens. Bottom Panel: QC contributions to $\mu_s^{ex}$ (Eq. 6.13) for methane in methanol as a function of the inner-shell radius, $\lambda_m$, for the methanol methyl groups. The corresponding radius for the hydroxyl groups of methanol is $\lambda_h = \lambda_m + 0.3$ Å. Symbols: $\mu_s^{ex} =$ solid triangles; $kT \ln x_0 =$ solid circles; $-kT \ln p_0 =$ open circles; $\langle \epsilon \rangle =$ solid squares; $\beta \sigma^2 / 2 =$ open squares. Lines are drawn connecting the data points to guide the eye.
shell grows in size. The outer-shell contribution to $\mu_s^{ex}$, consisting of the mean binding energy plus the binding-energy fluctuation contribution, is also insensitive to the defined inner shell. Plots of the conditional binding energy probability distributions as a function of the inner-shell radii are given in the Appendix B.

From Eq. 6.15, the corresponding QC contributions to $\Gamma_{sc}$ are derived from the methanol concentration dependence of the free energy contributions to $\mu_s^{ex}$. This concentration dependence is remarkably linear for each QC contribution over the range...
Table 6.3: Methane-methanol preferential interaction parameter $\Gamma_{sc}$ (Eq. 6.15) in aqueous methanol solutions as a function of methanol mole fraction, calculated from the slope of $\mu_s^{ex}(\rho_c)$ in Fig. 6.5 compared with $\Gamma_{sc}$ obtained from KB theory (Eq. 6.2) using the methane-water and methane-methanol rdfs with a cutoff radius of 6.0 Å. Units for $\Gamma_{sc}$ are number of methanol molecules/methane molecule.

<table>
<thead>
<tr>
<th>methanol mole frac.</th>
<th>$\rho_c \times 10^3$ (mol/cm$^3$)</th>
<th>$\Gamma_{sc}$ (Eq. 6.15)</th>
<th>$\Gamma_{sc}$ (Eq. 6.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>2.64</td>
<td>0.25</td>
<td>0.32</td>
</tr>
<tr>
<td>0.10</td>
<td>4.98</td>
<td>0.47</td>
<td>0.61</td>
</tr>
<tr>
<td>0.20</td>
<td>8.94</td>
<td>0.84</td>
<td>1.22</td>
</tr>
<tr>
<td>0.30</td>
<td>12.14</td>
<td>1.14</td>
<td>1.48</td>
</tr>
</tbody>
</table>

of concentrations considered, as shown in Fig. 6.5, which indicates that $\Gamma_{sc}/\rho_c$ is independent of the co-solvent concentration. $\Gamma_{sc}$ as a function of methanol concentration, calculated using Eq. 6.15 is given in Table 6.3 and compared with $\Gamma_{sc}$ calculated using the KB theory expression, Eq. 6.2 with the methane-water and methane-methanol rdfs in Figs. 6.1 and 6.2.

The comparison reveals that a consistently smaller $\Gamma_{sc}$ is derived from QC theory. For the 10 mol% methanol solution, the difference of -0.14 corresponds almost exactly to the outer-shell QC contribution, -0.15, for the defined inner shell used in the calculations (see $\lambda_w = 3.3$ Å in Table 6.4). Of course, this outer-shell contribution will depend on the definition of the inner shell, so the almost exact correspondence found for $\lambda_w = 3.3$ Å is fortuitous. Nonetheless, the results in Table 6.4 indicate that the outer-shell contribution lowers the QC-derived $\Gamma_{sc}$ regardless of the inner-shell specification. We note further that the inner-shell radii considered in Table 6.4 are also well below the cutoff radius, $R = 6.0$ Å, used to evaluate the KB integrals for $\Gamma_{sc}$ in Table 6.3. This observation is discussed below.
Table 6.4: QC contributions to the methane-methanol preferential interaction parameter $\Gamma_{sc}$ for 10 mol% methanol/water solution as a function of the inner-shell radius for water, $\lambda_w$. The corresponding radii for the methyl and hydroxyl groups of methanol are $\lambda_m = \lambda_w + 0.3$ Å, and $\lambda_h = \lambda_w + 0.6$ Å, respectively. Units for $\Gamma_{sc}$ are number of methanol molecules/methane molecule.

<table>
<thead>
<tr>
<th>$\lambda_w$ (Å)</th>
<th>$-kT \ln p_0$</th>
<th>$kT \ln x_0$</th>
<th>$\beta \langle \epsilon \rangle$</th>
<th>$-\beta \sigma^2/2$</th>
<th>$\Gamma_{sc}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>+0.38</td>
<td>+0.24</td>
<td>-0.12</td>
<td>-0.03</td>
<td>+0.47</td>
</tr>
<tr>
<td>3.4</td>
<td>+0.45</td>
<td>+0.23</td>
<td>-0.17</td>
<td>-0.03</td>
<td>+0.48</td>
</tr>
<tr>
<td>3.5</td>
<td>+0.52</td>
<td>+0.20</td>
<td>-0.21</td>
<td>-0.03</td>
<td>+0.48</td>
</tr>
<tr>
<td>3.6</td>
<td>+0.61</td>
<td>+0.14</td>
<td>-0.24</td>
<td>-0.03</td>
<td>+0.48</td>
</tr>
<tr>
<td>3.7</td>
<td>+0.72</td>
<td>+0.05</td>
<td>-0.26</td>
<td>-0.02</td>
<td>+0.49</td>
</tr>
<tr>
<td>3.8</td>
<td>+0.79</td>
<td>-0.05</td>
<td>-0.24</td>
<td>-0.01</td>
<td>+0.49</td>
</tr>
</tbody>
</table>

Figure 6.6: Methane-methanol preferential interaction parameter $\Gamma_{sc}$, and the QC contributions to $\Gamma_{sc}$ (Eq. 6.13) for 10 mol% methanol/water solution as the function of inner-shell radius, $\lambda_w$, for water. The corresponding radii for the methyl and hydroxyl groups of methanol are defined as $\lambda_m = \lambda_w + 0.3$ Å and $\lambda_h = \lambda_w + 0.6$ Å, respectively. Symbols: $\Gamma_{sc} =$ diamonds; QC contributions from $kT \ln x_0 =$ solid circles, $-kT \ln p_0 =$ open circles, $\langle \epsilon \rangle =$ solid squares, and $\beta \sigma^2/2 =$ open squares. Lines are drawn connecting the data points to guide the eye.
The impact of the inner-shell definition on the QC contributions to $\Gamma_{sc}$ for the 10 mol% methanol/water solution is shown in Table 6.4 and Fig 6.6. Inner-shell contributions from the free energy of cavity formation and chemical association are the dominant factors in obtaining $\Gamma_{sc} > 0$ for the smallest inner shell considered ($\lambda_w = 3.3 \, \text{Å}$). This contribution is, however, partially offset by a negative outer-shell contribution, as noted above.

As the defined inner shell grows in size, preferential interactions between methanol and methane within the inner shell are enhanced simply because larger inner shells can accommodate more of the larger co-solvent molecules. These enhanced preferential interactions are reflected in the increasingly dominant, positive contribution from the free energy of cavity of formation (solvent molecular packing). The additional co-solvent molecules within the larger inner shells do not contribute to these preferential interactions through enhanced association with the solute, as seen in the diminishing contribution from the free energy of chemical association. In fact, this contribution approaches zero for $3.7 \, \text{Å} < \lambda_w < 3.8 \, \text{Å}$, when the most probable number of methanol molecules within the defined inner shell is one.

The negative outer-shell contribution grows in magnitude as the defined inner shell is expanded, which reduces $\Gamma_{sc} > 0$. This contribution compensates for the positive inner-shell contribution that enhances $\Gamma_{sc} > 0$ as the inner shell becomes larger, and reflects a thermodynamic weighting that favors inner-shell molecular clusters with fewer rather than more methanol molecules. The importance of this outer-shell contribution is seen especially in Table 6.4, where $\Gamma_{sc}$ is essentially independent of the inner-shell definition when the outer-shell contribution included.
Recall that the inner-shell radii for water and methanol around a central methane molecule are defined as the radial distances at which the rdfs in Figs. 6.1 and 6.2 first equal 1.0. These radii, by definition, include very little of the first solvation shell of methane. It is striking that $\Gamma_{sc}$ is reliably evaluated using such small proximal volumes around the solute. The defined inner-shell radii are also well below the cutoff radius, $R$, needed to accurately evaluate the KB integrals using Eq. 6.4. This point is emphasized in Fig. 6.7 where $\Gamma_{sc}$ derived from KB theory is plotted as a function of $R$ for the aqueous methanol solutions considered. The shaded region shows the range of inner-shell radii used to derive $\Gamma_{sc}$ from QC theory. Since only a small fraction of the first solvation shell around methane is included for these values of the cutoff radius, $\Gamma_{sc}$ is meaningless. $\Gamma_{sc} < 0$ obtained in this range simply reflects the slightly larger excluded volume for methanol molecules relative to water molecules. Reasonable convergence is obtained using a significantly larger cutoff radius of 8-10 Å for the lowest concentration of 5 mol% methanol, but oscillations in $\Gamma_{sc}$ persist up to 12 Å at higher methanol concentrations, and a lack of convergence is evident at the highest concentration of 30 mol% methanol where $\Gamma_{sc}$ begins to tail off at $R \sim 10$ Å.

It is important to note that the thermodynamic weighting of inner-shell molecular clusters of solvent with the solute in QC theory does not simply require the material balances for water and co-solvent between inner and outer shells to be satisfied, which has been shown to improve convergence in evaluating KB integrals from molecular simulations of closed systems [184]. Instead, the outer-shell contribution in QC theory evaluates the thermodynamic cost of forming solute-solvent molecular clusters that are accessible to the inner shell for a given inner-shell specification.
Figure 6.7: Preferential interaction parameter $\Gamma_{sc}$ for methanol with methane in aqueous methanol solutions at different methanol concentrations derived from KB theory as a function of the cutoff radius, $R$ (Eq. 6.4). Arrows on the right denote $\Gamma_{sc}$ at different methanol concentrations derived from the QC/PDT analysis (Table 6.3). The shaded region shows the range of inner-shell radii considered in that analysis. Methanol mole fractions are: 0.05 (solid line), 0.10 (dotted line), 0.20 (dashed line), 0.30 (dash-dot line).
This direct thermodynamic assessment of probabilities of finding certain local solvent densities around the solute, which is central to the accurate evaluation of $\Gamma_{sc}$ from QC theory, is implicit in KB theory where the difference in local water and co-solvent densities must be normalized by the difference in water and co-solvent densities unaffected by the presence of the solute – i.e., bulk densities for water and the co-solvent. This normalization is achieved only for relatively large simulation systems that permit the observation volume around the solute to encompass multiple solvation shells for both water and the co-solvent. In addition, QC theory opens the possibility of using approximate models, rather than detailed molecular configurations from large simulation systems, to evaluate this outer shell contribution to $\Gamma_{sc}$, which is likely to be another potential advantage in evaluating co-solvent preferential interactions in the vicinity of specific solvation sites for complex solutes.

6.5 Conclusions

The analysis of preferential interactions in aqueous solutions presented here based on QC theory defines important local and non-local contributions to the preferential interaction parameter, which quantifies these intermolecular interactions between the co-solvent and solute. Local co-solvent preferential interactions are described by the chemical association of water and co-solvent with the solute and the molecular packing of water and co-solvent within a defined inner shell around the solute. The non-local or outer-shell contribution to these preferential interactions arise from solute interactions with water and co-solvent outside the defined inner shell. The significance of this outer-shell contribution is to assign thermodynamic weights to the molecular clusters that can form between the solute and solvent (water and co-solvent) within the defined
inner shell, which accounts for the dependence of the inner-shell composition on the
definition of the inner shell. Unlike KB theory, which describes preferential molecular
interactions based on local water and co-solvent densities around the solute, QC
theory requires the evaluation of both the inner-shell and outer-shell contributions
for an accurate description of the solute-co-solvent preferential interaction parameter.

A key finding of this study is the preferential interaction parameter derived from
QC theory is independent of the definition of the inner shell when both these contribu-
tions are taken into account. In addition, an accurate determination of this parameter
is possible using inner-shell volumes that are significantly smaller than those required
for the accurate evaluation of KB integrals. Thus, local preferential interaction pa-
rameters can be derived for specific sites on topologically and chemically complex
macromolecular solutes, such as globular proteins. For proteins, hydration and co-
solvent preferential interactions at specific sites are a primary interest, and would be
more informative than a preferential interaction parameter averaged over the entire
protein-solvent interface. An attempt to characterize local preferential interactions
for protein solvation based on evaluating KB integrals found the local protein-solvent
interactions to be strongly influenced by the topology and the chemical composition
of sites neighboring solute site of interest [188]. QC theory also offers the possibility of
obtaining local preferential interaction parameters using approximate models, rather
than detailed molecular configurations from simulation trajectories, to compute inter-
molecular interactions with the outer-shell solution, which may be another attractive
advantage of this approach applied to protein solvation. An accurate estimation of
this outer-shell contribution from approximate solution thermodynamic models or
directly from the molecular simulations also circumvents the convergence problems
that are typically encountered in estimating preferential interaction parameter from KB integrals.
CHAPTER 7

Site Specific Co-solvent Preferential Interactions on Protein Surface through Quasi-Chemical Theory

7.1 Introduction

Protein solubility and stability in aqueous solutions can be manipulated by addition of a co-solvent into the solution [13, 155, 26]. The choice of co-solvent is a key step in the application of co-solvents for protein phase separation in food [146, 147, 148, 149] and pharmaceutical [150, 151, 26, 152, 153, 154] industries. For example, inorganic salts influence protein solubility depending on the position of its constituent ions in the Hofmeister series [26, 209]. The nature of the salt also affects crystallization kinetics and the crystal morphology [209, 15, 210]. Organic co-solvents can stabilize the secondary structure or destabilize the native structure, leading to protein aggregation [174, 175]. Glycerol stabilizes the native structure of proteins and is often used as a co-solvent in protein crystallization to enhance the growth of protein crystals [155, 26]. While urea and guanidium chloride are the widely used as co-solvents for protein-denaturation [26, 166] in protein expression and purification processes.
Co-solvent effects on the solubility of proteins in aqueous solution are described through co-solvent preferential interactions with the protein [24, 26, 25]. The preferential interaction parameter, $\Gamma_{pc}$, quantifies preferential interactions as the extent of co-solvent and water partitioning in the vicinity of protein surface relative to bulk solution. For $\Gamma_{pc} > 0$, the local density of co-solvent near the protein surface is greater than that in bulk solution, while for $\Gamma_{pc} < 0$ the local density of co-solvent near the protein is less than that in bulk solution. $\Gamma_{pc}$ can be measured using equilibrium dialysis [162], high precision densitometry [158, 159, 160] and vapor pressure osmometry [157].

7.2 Theory

Kirkwood-Buff (KB) solution theory provides a rigorous statistical thermodynamic framework for determining the preferential interaction parameter [31, 32] from molecular distribution functions. For the preferential interaction parameter of a co-solvent (subscript $c$) with protein (subscript $p$), KB theory gives,

$$\Gamma_{pc} = 4\pi \rho_c \int_{0}^{\infty} (g_{pc}(r|\mu VT) - g_{pw}(r|\mu VT)) r^2 dr ,$$

(7.1)

where, $g_{pc}(r|\mu VT)$ and $g_{pw}(r|\mu VT)$ are co-solvent-protein and water-protein radial distribution functions (rdfs), respectively, in an open system. Simulation studies at constant NVT or NPT approximate the integral in Eq. 7.1 by defining a spherical subsystem of radius, $R$, within the larger simulation system where $R$ is the distance from the center of the closest surface atom of protein,

$$\Gamma_{pc}(R) = 4\pi \rho_c \int_{0}^{R} (g_{pc}(r) - g_{pw}(r)) r^2 dr .$$

(7.2)
Preferential interactions parameters are conveniently determined by simply counting
the number of co-solvent ($n_c$) and water ($n_w$) in the vicinity of the protein in excess
of that observed in bulk solution \[163, 164, 165, 166, 167\],
\[
\Gamma_{pc}(R) = n_c(R) - \frac{(N_c - n_c(R))}{(N_w - n_w(R))} n_w(R),
\] (7.3)
and have been reported to be invariant beyond 6-8 Å from the protein surface.

The local preferential interaction parameters for the constituent amino acids on a
protein surface are found to be strongly influenced by the topology and neighboring
amino acid composition \[164\]. Hence, the local interactions should be evaluated at
molecular level accounting for the topology and chemical heterogeneity of protein
surface. The co-solvent preferential interaction parameters for molecular solutes,
however, are sensitive to the upper limit of integration in Eq. 7.2 and converges only
at very large R, requiring inordinately large simulations for reliable estimation \[168,
169, 170, 171\]. The convergence of the integrals in KB theory can be improved
by dampening the long-range oscillations in solute-co-solvent and solute-water rdfs
through smearing, which assumes multiple reference points for a solvent molecule
instead of a single reference point \(\text{e.g.}, \) center of mass \[34, 35, 134, 211\].

Recently we adopted an entirely different approach for evaluating co-solvent pref-
erential interactions based on quasi-chemical (QC) theory and the potential distribu-
tion theorem \[212\]. We start with the following definition of the preferential interac-
tion parameter, expressed as a partial derivative of $\beta \mu_p^{ex}$ \[212\],
\[
\Gamma_{pc} = -\rho_c \left( \frac{\partial \beta \mu_p^{ex}}{\partial \rho_c} \right)_{T,P,\rho_p},
\] (7.4)
where the excess chemical potential of protein in aqueous solution is defined [20, 213, 214] as,

$$\beta \mu_p^{ex} = \ln x_0 - \ln p_0 + \ln \langle e^{+\beta\epsilon} | n, m = 0 \rangle .$$  \hspace{1cm} (7.5)

$x_0$ is the probability of finding no water or co-solvents molecules within a defined proximal volume or inner shell around the protein; $p_0$ is the probability of finding a cavity in bulk solution corresponding to the inner shell; and $\ln \langle e^{+\beta\epsilon} | n, m = 0 \rangle$ is the outer shell contribution that accounts for the interaction of protein with the solvent in the outer shell under the condition that there is no solvent in the inner shell. If the probability distribution of interaction energies, $\epsilon$, is accurately gaussian [113, 202, 203, 136], the excess chemical potential adopts a simple form in terms of the mean, $\langle \epsilon \rangle$, and variance, $\sigma^2$, of the distribution of interaction energies,

$$\beta \mu_p^{ex} = \ln x_0 - \ln p_0 + \beta \langle \epsilon \rangle + \beta^2 \sigma^2 / 2 .$$  \hspace{1cm} (7.6)

Evaluating the partial derivative in Eq. 7.4 gives,

$$\Gamma_{pc} = -\rho_c \left( \frac{\partial \ln x_0}{\partial \rho_c} \right)_{T,P,\rho_p} + \rho_c \left( \frac{\partial \ln p_0}{\partial \rho_c} \right)_{T,P,\rho_p} - \rho_c \left( \frac{\partial \ln \langle e^{+\beta\epsilon} | n, m = 0 \rangle}{\partial \rho_c} \right)_{T,P,\rho_p} .$$  \hspace{1cm} (7.7)

The first two terms on the right side of this equation describe inner-shell preferential interactions in terms of the effect of co-solvent bulk concentration on the composition of the solvent molecular clusters that associate with the solute in the inner shell. The rightmost term describes the preferential stabilization of certain solvent molecular clusters over others in the inner shell depending on the bulk concentration of co-solvent. The assessment of the effect of bulk co-solvent concentration in finding a certain solvent molecular cluster in the inner shell, and on the thermodynamic cost
for forming this solvent molecular cluster, \( i.e., \) the outer-shell contribution, enables an accurate estimation of \( \Gamma_{pc} \) using inner shell volumes significantly smaller than required for KB theory \([212]\). The convenience of choosing small inner shell volumes allows defining an inner shell proximal to a specific site on large protein surfaces to determine local preferential interaction parameter for the site.

7.3 Methods

7.3.1 Simulation Details

All molecular dynamics (MD) simulations were carried out at constant NPT using NAMD 2.6 \([58]\) and the CHARMM27 force field \([139]\). Lysozyme (PDB ID: 1LYZ) \([55]\) was solvated in a cubic box of water or well equilibrated aqueous methanol/urea solution, after removing all water and co-solvent molecules within 4 Å of any protein heavy atom. Lysozyme in pure water simulation comprised of 6727 water molecules. For lysozyme in 2 mol\% urea solution, the protein was solvated by 180 urea molecules and 8820 water molecules, while it was solvated by 900 methanol molecules and 8100 water molecules in 10 mol\% aqueous methanol solution. The simulations of neat water and of bulk urea and methanol solutions were also carried with the same number of co-solvent and water molecules, but in the absence of the protein. Water was modeled with TIP3P potential \([140]\), while the parameters for methanol and urea were taken from the recently published force fields parameters \([215, 216]\).

Each system was initially minimized for 20,000 steps by conjugate-gradient minimization, and then equilibrated for 3 ns at a constant temperature of 300 K and 1 bar pressure. A time step of 2 fs was used and configurations were saved every 0.1 ps over a production run of 20 ns. The protein was held rigid during the simulation.
Table 7.1: Protein sites of interest for local co-solvent preferential interaction analysis

<table>
<thead>
<tr>
<th>Protein site index</th>
<th>residue number</th>
<th>residue name</th>
<th>molecular group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14</td>
<td>Arg</td>
<td>N°1H₂</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>Arg</td>
<td>N°2H₂</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>Arg</td>
<td>C=O&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>4</td>
<td>127</td>
<td>Cys</td>
<td>C=O&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>5</td>
<td>128</td>
<td>Arg</td>
<td>N°1H₂</td>
</tr>
<tr>
<td>6</td>
<td>128</td>
<td>Arg</td>
<td>N°2H₂</td>
</tr>
<tr>
<td>7</td>
<td>129</td>
<td>Leu</td>
<td>C=O</td>
</tr>
</tbody>
</table>

<sup>a</sup> carbonyl carbon of these groups are not solvent accessible, therefore, the proximal volume and the inner shell volume around these groups are estimated based on only oxygen atom of the group.

The positions of all hydrogen atoms in the system were constrained using the SHAKE algorithm [60].

Temperature was held constant at 300 K by applying Langevin dynamics to all heavy atoms using a damping coefficient of 1 ps<sup>−1</sup>. A constant pressure of 1 bar was maintained using a Nosé-Hoover Langevin piston [141, 142] with a period of 200 fs and a decay of 100 fs. Periodic boundary conditions were imposed, and the particle-mesh Ewald method [143] with a real-space cutoff of 12 Å was used in computing the electrostatic interactions. Lennard-Jones interactions were smoothly switched to zero at 12 Å starting from 10 Å.

7.3.2 Analysis

We evaluated urea and methanol co-solvent preferential interactions at seven solvent accessible molecular sites (Table. 7.1) on the protein lysozyme. Fig. 7.1 shows these molecular sites colored according to their atom-type. The other heavy atoms of the protein are colored in orange. X-ray crystallography of lysozyme crystals grown
from aqueous urea solution show that urea binds to Arg 14 in N-terminus, and Cys 127, Arg 128 and Leu 129 in C-terminus of the protein [217]. Arg14 O, Arg128 N\textsuperscript{\eta2} and Leu129 O are also crystal contact points for tetragonal crystal structure of lysozyme [218].

The proximal solvent accessible region for a protein site is evaluated as the sum of proximal sub-domains for its constituent heavy atoms on protein surface. For a heavy atom at $\vec{r}_\alpha$ on the surface, the proximal sub-domain is determined from the criterion,

$$|\vec{r} - \vec{r}_\alpha| - \sigma_\alpha = \text{min}_{\beta=1,\ldots,n} |\vec{r} - \vec{r}_\beta| - \sigma_\beta,$$

(7.8) where $n$ is the total number of heavy atoms on the protein surface. Protein atoms with non-zero solvent excluded surface area [219] are identified as surface atoms, while other atoms are considered as buried atoms. The size disparities among protein
atoms are taken into account in Eq. 7.8 by accounting for their solvent excluded-volume radius, $\sigma_\beta$. The water excluded-volume radii are given in Table. 7.2.

Using Eq. 7.8, the number of water and co-solvent (center of mass) molecules proximal to a heavy atom of a protein site are counted as a function of their distance from the heavy atom. To evaluate the proximal solvent density, the proximal volume for the heavy atom at the given radial distance is computed as the fractional volume of spherical shell – corresponding to the spherical bin size for density calculation – such that any point in this volume is proximal to the heavy atom of interest. The proximal solvent radial distribution function around the heavy atom is normalized to the bulk solvent density.

We smear the solvent and co-solvent positions in calculating the proximal radial distribution functions to improve convergence of integrals in KB theory. To smear the proximal radial distribution functions, multiple solvent reference centers within a spherical volume around each heavy atom of the solvent molecules are assumed to be of equal probability \[211, 34\], instead of a single reference point such as solvent center of mass. The different smearing radii for the different heavy atoms of co-solvent and water molecules were used based on their molecular sizes: oxygen – 2.3 Å, nitrogen – 2.5 Å, and carbon – 3.0 Å.

The inner shells proximal to these protein sites for QC analysis are defined by a set of conditioning radii for every heavy atom of the co-solvent/water molecules. The conditioning radii for each heavy atom of urea and methanol can be related to the water conditioning radius, $\lambda_w$, by $\lambda_i = \lambda_w - 1.3 + R_i$ \[220\], where $R_i$ is one half of the distance at which radial distribution function $g_{ii}(r)$ in bulk co-solvent solution first reaches unity. The $R_i$ for the heavy atoms of urea and methanol are given in
Table 7.2: Classification of molecular groups on protein surface with their water excluded-volume radii, identified as the smallest radial distance at which the proximal radial distribution of water around these groups reach unity.

<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
<th>$\sigma$ (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>–N–</td>
<td>amine and other nitrogen groups</td>
<td>2.74</td>
</tr>
<tr>
<td>O=(C)</td>
<td>carbonyl oxygen group</td>
<td>2.52</td>
</tr>
<tr>
<td>OH</td>
<td>hydroxyl oxygen group</td>
<td>2.61</td>
</tr>
<tr>
<td>–S–</td>
<td>sulphur group</td>
<td>3.15</td>
</tr>
<tr>
<td>C=(O)</td>
<td>carbonyl carbon group</td>
<td>3.09</td>
</tr>
<tr>
<td>ACH or AC–(C)</td>
<td>aromatic carbon group including those connected to a carbon group</td>
<td>3.15</td>
</tr>
<tr>
<td>AC–(O/N)</td>
<td>aromatic carbon connected to a polar oxygen/nitrogen group</td>
<td>2.96</td>
</tr>
<tr>
<td>CH</td>
<td>methylene (CH) group</td>
<td>3.31</td>
</tr>
<tr>
<td>CH$_{2/3}$–(O/N)</td>
<td>methyl or methylene group connected to a polar oxygen/nitrogen group</td>
<td>3.24</td>
</tr>
<tr>
<td>CH$_2$–(C)</td>
<td>methylene group connected to a carbon group</td>
<td>3.29</td>
</tr>
<tr>
<td>CH$_3$–(C)</td>
<td>methyl group connected to a carbon group</td>
<td>3.26</td>
</tr>
</tbody>
</table>
Table 7.3: $R_i$ for the heavy atoms of urea and methanol, defined as one half of the distance at which radial distribution function $g_{ii}(r)$ in bulk co-solvent solution first reaches unity.

<table>
<thead>
<tr>
<th>Solvent heavy atom, $i$</th>
<th>$R_i$ (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>urea</td>
<td></td>
</tr>
<tr>
<td>carbonyl carbon</td>
<td>1.95</td>
</tr>
<tr>
<td>carbonyl oxygen</td>
<td>1.80</td>
</tr>
<tr>
<td>amino nitrogen</td>
<td>1.65</td>
</tr>
<tr>
<td>methanol</td>
<td></td>
</tr>
<tr>
<td>methyl carbon</td>
<td>1.80</td>
</tr>
<tr>
<td>hydroxyl oxygen</td>
<td>1.40</td>
</tr>
</tbody>
</table>

Table 7.3. The inner shell occupancies of water and co-solvent molecules are counted such that at least one heavy atom of a solvent molecule is within the defined inner shell volume. The probability of finding no water and co-solvent within the inner shell around a protein site provides the required $x_0$ for this site.

Evaluation of $p_0$ for a protein site requires identification of different co-solvent-water molecular clusters in bulk co-solvent solution. The occupancies of co-solvent and water molecules in a volume corresponding to the inner shell of a protein site are computed from the simulations in absence of the protein, as described below. In order to accurately define the irregular inner shell volume around a protein site, the structure of the entire protein molecule is stenciled within the simulation shell and the domain proximal to the protein site is determined (Eq. 7.8). This proximal volume is larger than the volume corresponding to the inner shell for the protein site, as it also volume not accessible to the solvent due to the presence of inner buried protein atoms. The solvent inaccessible volume is modeled as the vdW volume of all buried atoms. The buried protein heavy atoms were assigned vdW radii based on the
atom-type: carbon – 1.7 Å, nitrogen – 1.55 Å, oxygen – 1.52 Å and sulphur – 1.80 Å.

Therefore, solvent occupying the inner shell volume of a protein site are counted as those solvent molecules having at least one heavy atom within the proximal region around a stenciled protein site but are not within the volume corresponding the vDW volume of buried atoms of the protein. The fraction of solvent configurations in which no water and co-solvent molecule is found in the inner shell volume of a protein site gives $p_0$ for the site.

The outer shell contribution involves calculating the interaction energy of a protein site with water and co-solvent in the outer shell. This interaction energy is computed as the sum of electrostatics and van der Waals interactions of the protein site with the outer shell water and co-solvent molecules, under the condition of no water or co-solvent molecules within the defined inner shell. The interactions of a protein site with other protein atoms or with its periodic images in the adjacent simulation cells are not included. The electrostatic interactions were computed using generalized reaction field [107, 108, 109, 110] with a cut-off radius of 12 Å and the dispersion interactions were also truncated at the same cut-off. The dielectric constant, $\epsilon_s$, used in the energy calculations are 104 for pure water, 93 for 2 mol% of aqueous urea solution and 78 for 10 mol% of aqueous methanol solution. Dielectric constants of the solvent were computed from the observed fluctuations [221] in dipole moment in simulation of aqueous co-solvent solution in absence of the protein molecule –

$$\epsilon_s = 1 + \frac{\langle M^2/V \rangle}{3kT}, \quad (7.9)$$

where the dipole moment, $M = |\sum_i q_i r_i|$, is defined as the magnitude of net dipole vector for a given solvent configuration computed as the summation of product of
Figure 7.2: Co-solvent preferential interaction parameter of lysozyme in aqueous solution of 2 mol% urea (solid line) and 10 mol% methanol (dashed line), evaluated by counting the number of co-solvent (center of mass) and water (oxygen) molecules in the vicinity of the entire protein surface (Eq. 7.3).

charge ($q_i$) and position ($r_i$) of every atom $i$ in the system, and $V$ is the volume of the system at that configuration.

The conditional probability distributions of interaction energies for all protein sites are approximately gaussian (refer to Fig. C.1 of Appendix C). The outer shell contribution was calculated as the sum of mean interaction energy ($\langle \epsilon \rangle$) and variance ($\sigma^2$) of the distribution of interaction energies. The site-specific preferential interaction parameters were evaluated as the derivative of each QC contribution to excess chemical potential (Eq. 7.7), computed as the difference between the corresponding values in the aqueous co-solvent solution and in pure water.
7.4 Results and Discussion

Fig. 7.2 shows co-solvent preferential interaction parameters for urea and methanol with lysozyme averaged over the entire surface of the protein, computed by counting the number of urea or methanol molecules and water molecules in the vicinity of lysozyme in 2 mol% urea and 10 mol% methanol solutions (Eq. 7.3) as a function of the distance from the center of the closest heavy atoms on lysozyme surface. The calculated $\Gamma_{pc}$ in either solution converges at $\sim 10$ Å. On an average, lysozyme is preferentially solvated ($\Gamma_{pc} > 0$) by urea and methanol, but to a much greater extent by urea. For urea, $\Gamma_{pc}$ is 11.6 indicating that about 11–12 urea molecules are found in excess in the vicinity of lysozyme than that found in bulk solution. While in 10 mol% methanol solution, $\Gamma_{pc}$ is 5.2, corresponding to the presence of 5 methanol molecules in the vicinity of lysozyme in excess to that in bulk solution. The preferential interaction parameter of urea with lysozyme have been reported to depend on the urea force field used in the study [167, 166]. For the urea force field [216] and concentration considered in this study, Shukla et al., [167] found the urea preferential interaction of lysozyme to be only 6.6. Different lysozyme pdb structures are used in these two studies. Moreover, the protein in this work was fixed during the simulation, as our interest is in identifying urea binding at the crystals contacts, while it was not fixed in the other study.

Fig. 7.3 shows the local preferential interaction parameters for lysozyme sites listed in Table. 7.1 computed from the smeared proximal rdfs of urea or methanol and water molecules around the sites. For urea, $\Gamma_{ac} \geq 0.0$ at all sites, indicating that the protein sites are either preferentially solvated by urea (sites 1, 2, 3, 5, 6, 7) or neutral to urea (site 4, magnitude of $\Gamma_{ac} \leq 0.05$). The convergence of $\Gamma_{ac}$ for protein
Figure 7.3: Local co-solvent preferential interaction parameters of the protein sites in aqueous solution of 2 mol% urea (top panel) and 10 mol% methanol (bottom panel) evaluated by KB theory using smeared proximal radial distribution functions of water (oxygen) and co-solvent (local smearing around each heavy atoms) molecules around the protein sites. Legends: solid line = site 1; dashed line = site 2; dotted line = site 3; long dashed line = site 4; dash dotted line = site 5; long dash dotted line = site 6; dash dot dotted line = site 7.
Figure 7.4: Local co-solvent preferential interaction parameters of the protein sites in aqueous solution of 2 mol% urea (top panel) and 10 mol% methanol (bottom panel) evaluated using QC theory (Eq. 7.7) as a function of the conditioning radius, $\lambda_w$, for water oxygens defining the inner shell. Legends: solid circle = site 1; open circle = site 2; solid square = site 3; open square = site 4; solid triangle up = site 5; open triangle up = site 6; open triangle down = site 7. Lines are drawn connecting the data points to guide the eye.
Figure 7.5: Comparison of the local co-solvent preferential interaction parameters of the protein sites in aqueous solution of 2 mol% urea (top panel) and 10 mol% methanol (bottom panel) evaluated using QC theory (Eq. 7.7), for an inner shell defined by water conditioning radius of $\lambda_w = 2.8 \, \text{Å}$, against its inner shell contribution and the values evaluated by KB theory using smeared radial distribution functions of water and co-solvent around the sites.
site 6 is observed at $R > 12 \, \text{Å}$, while for other proteins $\Gamma_{ac}$ converges at $\sim 10 \, \text{Å}$. The convergence is delayed for the site 6, as more urea molecules are found proximal to this most solvent accessible protein site. In methanol solution, 3 proteins sites – site 1, 2 and 7 – are preferentially hydrated ($\Gamma_{ac} < 0.0$), while for the other 4 protein sites $\Gamma_{ac} \approx 0.0$, with a magnitude of $\leq 0.05$, indicating no preference to either methanol or water.

Fig. 7.4 shows the corresponding site-specific preferential interaction parameters computed using QC theory (Eq. 7.7). In urea solution $\Gamma_{ac} > 0$ for all protein sites, except the site 6 where $\Gamma_a \approx 0$. The variation in $\Gamma_{ac}$ as a function of inner shell volume defined by the conditioning radius for water, $\lambda_w$, is large for protein site 7. In case of methanol solution, $\Gamma_{ac} > 0$ for protein sites 1, 2, 5 and 6 (all amine groups), $\Gamma_{ac} < 0$ for protein sites 3 and 7 and $\Gamma_{ac} \approx 0$ for the protein site 6. The figures showing the individual QC contributions to $\Gamma_{ac}$ are given in the Appendix C. The contributions from outer shell interaction energy and the inner shell chemical association to the local preferential interactions parameters are dominant over the packing contribution, in-contrast to the result for methane in methanol solutions [212].

The variation in $\Gamma_{ac}$ as the function of the inner shell volume (Fig. 7.4) is largely due to observed variations in the variance of the conditional distribution of interaction energies (refer to Figs. C.5 and C.6 of Appendix C). The mean and standard deviation of the site-specific preferential interaction parameters calculated over different inner shell volumes considered in this study are given in Table. C.1 of the Appendix C. The largest standard deviation of 0.1 is observed for site 7, whose distribution of interaction energies are broader than other protein sites (Fig. C.1 of Appendix C).
The site-specific preferential interaction parameters evaluated by QC theory for an inner shell defined by $\lambda_w = 2.8$ Å are compared against those evaluated by KB theory at $R = 12$ Å in Fig. 7.5. It also shows QC inner shell contribution to the preferential interaction parameters at these sites. In urea solution, the values of the local preferential interaction parameter obtained from QC and KB theories are equivalent for protein sites 3 and 7 and are in close agreement for sites 1 and 2, but are widely different for the sites 4, 5 and 6. QC inner shell contribution to $\Gamma_{ac}$ at all protein sites are positive and are equal to or less than the value evaluated using KB theory. The protein site 6 is the most solvent accessible site and more urea molecules were found proximal to the site as indicated by $\Gamma_{ac}$ obtained from KB theory. However, $\Gamma_{ac}$ obtained from QC is $\approx 0$ as the chemical and packing contributions exactly counterbalanced each other, and the interaction energy of the protein site with the outer shell solvent medium showed no significant difference in urea solution and in pure water. Contrastingly, the number of urea molecules found proximal to the site 4 and 5 are smaller compared to site 6, but the interaction energies of these sites with the outside solvent medium is favorable in urea solution than that in water, hence, $\Gamma_{ac}$ obtained from QC theory is very large than that evaluated form KB theory.

In methanol solution, from KB analysis the protein sites 1, 2 and 7 are weakly preferentially hydrated ($-0.2 < \Gamma_{ac} < 0$), while other sites showed no preference to either methanol or water ($\Gamma_{ac} \approx 0$). Whereas the QC inner shell contribution to the preferential interactions suggest all the proteins sites are at least weakly preferentially solvated by methanol ($0 < \Gamma_{ac} < 0.2$). Furthermore, when the interaction energy of these site with the outside solvent medium is taken into account, it is found that
all the amine groups (sites 1, 2, 4 and 5) are preferentially solvated by methanol while the two carbonyl groups (sites 3 and 7) are preferentially hydrated and the other carbonyl group (site 4) showed no preference for either methanol or water. Considering the polar nature of the protein sites, it is not surprising to observe that interaction energies that include long range electrostatic interactions play a significant role in preferential interactions. In addition, unlike KB theory, QC analysis have also explicitly accounted for interaction of hydrogen atoms of the molecular group of the site with the solvent.

It is interesting that QC inner shell contributions to $\Gamma_{\alpha c}$, the first two terms in Eq. 7.7 at all proteins sites are positive in both urea and methanol solution. In either solution, both the terms are positive, however, the relative magnitude of contributions to $\Gamma_{\alpha c}$ are different (refer to Figs. C.5 and C.6 of Appendix C) suggesting different origins for favorable interactions. With addition of urea or methanol into water, at the investigated co-solvent concentration, the chemical association of solute with solvent increases ($x_0$ decreases) and the work reduced to create a cavity decreases ($p_0$ increases). But the former contribution towards the observed positive QC inner shell contribution to $\Gamma_{\alpha c}$ is larger in urea solution, while the latter contribution is larger in methanol solution. However, the outer shell contribution to overall $\Gamma_{\alpha c}$ is large and even it flips the nature of preferential interaction at some sites from that obtained from inner shell contribution.

### 7.5 Conclusions

Co-solvent preferential interactions at specific sites on protein surface is of primary importance to protein-protein interactions in aqueous solutions, and would be
more informative than a preferential interaction parameter averaged over the entire protein surface. QC theory provides the appropriate framework in determining local preferential interactions by the direct assessment of inner and outer shell contributions to the preferential interaction at the site. Through the estimation of smeared proximal radial distribution around the heavy atoms of the protein site, the local preferential interaction parameters of the sites are evaluated by KB theory, however, it does not explicitly account for the long-range nature of the electrostatic interactions and the interaction of the solvent medium with hydrogen atoms at the protein site. Considering the polar nature of investigated protein sites, it is necessary to account for the electrostatic interactions. QC theory clearly reveals the importance of these interactions as they contribute dominantly to the local preferential interactions of the protein sites. One protein site is neutral toward urea and water, except for that site, other proteins sites are found to be preferentially solvated by urea. In methanol solution, amino groups are solvated by methanol, while the carbonyl groups are either preferentially hydrated or showed no preference to either methanol or water.
CHAPTER 8

Effect of PEG End-Group Hydrophobicity on Lysozyme-Lysozyme Interactions

8.1 Introduction

Low molecular-weight poly(ethylene glycol) (PEG) is commonly used as an additive to induce protein precipitation and crystallization from aqueous solutions [9, 13, 36, 37, 38]. The molecular mechanism for PEG-induced protein phase separation is thought to be depletion attraction, an entropically driven process that produces an effective protein-protein attraction by preferentially excluding polymer from the interstitial regions between the protein molecules [222, 223, 37, 38, 224, 225]. In the primitive depletion-attraction model, polymer-protein interactions are considered to be purely repulsive, excluded volume interactions [226, 227, 228, 229, 230].

However, excluded volumes only qualitatively describe protein phase separation induced by low molecular-weight PEG [231, 232, 223, 233]. Attractive, energetic interactions in addition to entropic depletion have been implicated in analyzing small-angle neutron scattering measurements of bovine serum albumin-PEG osmotic second virial coefficients [234] where it was determined that protein-PEG attractions must be taken into account for an accurate, quantitative description of the virial coefficients. A short-range attraction that overcomes longer-range steric repulsion has also
been measured using the surface force apparatus for streptavidin and low molecular-weight, methyl-terminated PEG anchored to lipid bilayers [235]. The measurements were interpreted based on a mechanism that couples short-range energetic attractions and entropic depletion in the sense that conformational rearrangements of the polymer in the presence of the protein facilitate the formation of attractive contacts between exposed sites on the protein surface and the ethylene oxide segments of the polymer [235].

Light scattering studies of the influence of low molecular-weight PEG on the lysozyme potential of mean force similarly found that an accurate model of the measured osmotic virial coefficients must account for an energetic attraction between the protein and the polymer in addition to entropic depletion [236, 237]. Competition between PEG-lysozyme short-range attraction and long-range repulsion is observed in one study [237] as a non-monotonic dependence of the lysozyme potential of mean force on PEG concentration with an attractive well near the crossover concentration from dilute to semi-dilute solution behavior for the polymer.

The effect of PEG-lysozyme interactions on the lysozyme potential of mean force is investigated here by measuring osmotic virial coefficients for lysozyme in aqueous solutions with low molecular-weight, hydroxy-terminated (hPEG) and methyl-terminated (mPEG) poly(ethylene glycol). Previous studies have established that PEG exhibits at least some hydrophobic character [238, 239], and that aqueous solution behavior for low molecular-weight PEG is indeed sensitive to the hydrophobicity of the polymer end-groups [240]. The modification of end-group hydrophobicity for a given PEG molecular weight predominantly affects intermolecular interactions between the
polymer and protein, which allows us to probe the effect of these interactions on the lysozyme potential of mean force, independent of steric repulsion.

8.2 Theory

The Gibbs free energy of an aqueous solution of protein (component 1) and polymer (component 2) is expressed as a series expansions in terms of solute concentrations, \( c_1 \) and \( c_2 \):

\[
\frac{G}{kT} = c_1 \ln c_1 + c_2 \ln c_2 + B_{11}c_1^2 + 2B_{12}c_1c_2 + \ldots
\]

where \( B_{11} \) and \( B_{22} \) are the protein-protein and polymer-polymer osmotic second virial coefficients, respectively, \( B_{12} \) is the protein-polymer osmotic second virial coefficient, and \( C_{111} \) and \( C_{112} \), are osmotic third virial coefficients, related to three-body interactions among the protein and polymer molecules. The virial coefficients are obtained be measuring the excess Rayleigh ratio, \( R_{90} \), defined as the intensity of the light scattered from the solution in excess of light scattered from the pure solvent.

For an aqueous solution containing protein as the single solute, the ratio of the protein concentration to the excess Rayleigh ratio,

\[
\frac{Kc_1}{R_{90}} = \frac{1}{M_1} + 2B_{11}c_1,
\]

plotted as a function of protein concentration is linear (Zimm plot) with the slope related to the protein osmotic second virial coefficient and the intercept equal to the inverse of the protein molecular weight \([15, 241, 242]\). The optical constant in Eq. 8.2 is \( K \equiv 4\pi^2n_0^2(\partial n/\partial c_1)^2/\lambda^4N_A \), with \( n_0 \) the refractive index of the solvent, \( \lambda \) the wavelength of the scattering light and \( N_A \) Avogadro’s number.
For aqueous solutions containing PEG in addition to the protein, the measured quantity of interest is the difference between the excess Rayleigh ratio for this solution, \( R_{90}^{1+2} \), and the excess Rayleigh ratio for the aqueous solution containing only the polymer, \( R_{90}^2 \), \[242, 243, 236\],

\[
\frac{Kc_1}{(R_{90}^{1+2} - R_{90}^2)} = \frac{1}{M_{1\text{app}}} + 2B_{11\text{app}}^1 c_1. \tag{8.3}
\]

The slope and intercept of the corresponding Zimm plot are related to an apparent protein-protein osmotic second virial coefficient,

\[ B_{11\text{app}} = B_{11} + [3C_{112} - 2B_{12}^2 M_2 + 6mC_{112} - 4mB_{11}B_{12}M_1] c_2 , \tag{8.4} \]

and an apparent protein molecular weight,

\[
\frac{1}{M_{1\text{app}}} = \frac{1}{M_1} + 4mB_{12}c_2 , \tag{8.5}
\]

that depend on the polymer concentration \[236\]. Here \( m \equiv M_2(\partial n/\partial c_2)/M_1(\partial n/\partial c_1) \) with \( n \) the refractive index of the solution and \( M_i \) the molecular weight of component \( i \). The effective osmotic second virial coefficient,

\[ B_{11\text{eff}} = B_{11} + (3C_{112} - 2B_{12}^2 M_2) c_2 , \tag{8.6} \]

is obtained from Eq. \[8.4\] and defines the effect of polymer concentration on protein-protein interactions,

\[ \partial B_{11\text{eff}}/\partial c_2 = 3C_{112} - 2B_{12}^2 M_2 , \tag{8.7} \]

with the cross virial coefficients, \( B_{12} \) and \( C_{112} \), derived from the polymer concentration dependence of \( M_{1\text{app}}^\text{app} \) and \( B_{11\text{app}}^\text{app} \).
8.3 Materials and Methods

Hen egg white lysozyme (3× crystallized, Catalog No. L6876), hPEG (\(M_n = 1000, 3515\)), mPEG (\(M_n = 1000, 44589\)) were purchased from Sigma Aldrich and used without any further purification. The solution pH was maintained at 6.2 and 7.0 using sodium phosphate buffer (sodium phosphate monobasic monohydrate: S9638; sodium phosphate dibasic: S7907, Sigma Aldrich) and bis-Tris buffer (bis-Tris hydrochloride: B6032; bis-Tris: B9754, Sigma Aldrich), respectively. The ionic strength was adjusted to 0.01 M and 0.8 M (including buffer ions) using NaCl (Z273287, Sigma Aldrich)

All solutions were prepared with filtered deionized water using a Barnstead Nanopure Ultraviolet water filter system (Barnsted International). Buffer solutions were filtered using Millipore vacuum filtration unit with a 0.1 \(\mu\)m PVDF membrane. Stock solutions contained \(~4\) mg/ml of lysozyme at different PEG concentrations ranging from 0-110 mg/ml. To minimize the protein aggregation, the protein solutions were prepared less than two hours before each experiment. Prior to an experiment, the solutions were filtered using a 0.22 \(\mu\)m Millipore syringe top filter, and also centrifuged at 3500 g for 15 minutes to remove any microscopic air bubbles. All experiments were carried out at a constant temperature of 25°C.

Static light scattering data were collected at an angle of 90° using miniDawn Treos (Wyatt Technology, CA) equipped with Gallium Arsenide (GaAs) laser, operating at a wavelength of 658 nm. Protein concentrations were measured using Varion Prostar 325 UV spectrophotometer. UV absorption of lysozyme solutions was measured at 280 nm, and the protein concentration was determined using the extinction coefficient of 2.64 liter/g cm \([15, 23]\). The solution composition during the experiment was controlled using a composition gradient pump system (Calypso 1.2.3.6), equipped
with three 1 ml syringe-piston pumps, by adjusting the relative injection volume of PEG-lysozyme solution to the PEG solution contained in different syringe pumps. Before each experiment, the system was flushed with buffer solution for two hours, followed by flushing with the PEG solution for an additional half an hour until a constant baseline in the light scattering detector was observed. The refractive index increments \( \frac{\partial n}{\partial c_1} \) of 0.131 ml/g and 0.126 ml/g were measured for hPEG and mPEG, respectively, using Rudolph J57 Automatic Refractometer. A refractive index increment of 0.185 ml/g was used for lysozyme [230].
Figure 8.2: Zimm plots (Eq. 8.3) of light scattering from aqueous solutions containing lysozyme (1) and mPEG (2) as a function of lysozyme concentration at 6.2 pH and 0.8 M ionic strength (including 0.2 M sodium phosphate buffer and 0.5 M NaCl) relative to light scattering from mPEG solutions at the same solution conditions. Symbols correspond to mPEG concentrations (bottom to top) of 0, 50, 70, and 90 mg/ml. Lines are linear fits to the data. The data at each mPEG concentration, starting with 50 mg/ml, are shifted upwards by 2 units for visual clarity.
Figure 8.3: Zimm plots (Eq. 8.3) of light scattering from aqueous solutions containing lysozyme (1) and hPEG (2) as a function of lysozyme concentration, at 7.0 pH and 0.01 M ionic strength (including 10 mM Bis-Tris buffer) adjusted using NaCl, relative to light scattering from hPEG solutions at the same solution conditions. Symbols correspond to hPEG concentrations (bottom to top) of 0, 30, 50, 70, 90, and 110 mg/ml. Lines are linear fits to the data. The data at each hPEG concentration, starting with 30 mg/ml, are shifted upwards by 2 units for visual clarity.
Figure 8.4: Zimm plots (Eq. 8.3) of light scattering from aqueous solutions containing lysozyme (1) and mPEG (2) as a function of lysozyme concentration, at 7.0 pH and 0.01 M ionic strength (including 10 mM Bis-Tris buffer) adjusted using NaCl, relative to light scattering from mPEG solutions at the same solution conditions. Symbols correspond to mPEG concentrations (bottom to top) of 0, 30, 50, 70, 90, and 110 mg/ml. Lines are linear fits to the data. The data at each mPEG concentration, starting with 30 mg/ml, are shifted upwards by 2 units for visual clarity.
Figure 8.5: The dependence of the apparent molecular weight of the protein (Eq. 8.5) on polymer concentration, in hPEG (2) or mPEG (2)-lysozyme (1) aqueous solutions at ionic strengths of 0.8 M (top figures) and 0.01 M (bottom figures). Left panels: hPEG solutions; Right panels: mPEG solutions.
Figure 8.6: The dependence of the apparent osmotic virial coefficient (Eq. 8.4) of the protein on polymer concentration, in hPEG (2) or mPEG (2)-lysozyme (1) aqueous solutions at ionic strengths of 0.8 M (top figures) and 0.01 M (bottom figures). Left panels: hPEG solutions; Right panels: mPEG solutions.
8.4 Results and Discussion

Zimm plots at different constant concentrations of hPEG and mPEG are shown in Figs. 8.1 and 8.2 respectively, for aqueous lysozyme solutions at 6.2 pH and 0.8 M ionic strength. The $c_1 = 0$ intercept of the Zimm plot in the absence of PEG ($c_2 = 0$) in both figures corresponds to a lysozyme molecular weight of 13977 Da, in good agreement with the molecular weight calculated from the amino acid sequence: 14300 Da. The slope of this Zimm plot gives $B_{11} = -7.85 \times 10^{-4}$ mol ml/g$^2$, in agreement with $B_{11} = -4.95 \times 10^{-4}$ mol ml/g$^2$, reported previously for the same pH, ionic strength, and sodium phosphate buffer [236]. See Table 8.1 for a compilation of all osmotic virial coefficients. The negative $B_{11}$ indicates attractive lysozyme-lysozyme interactions, which is expected at high ionic strength.

The corresponding Zimm plots at 7.0 pH and 0.01 M ionic strength are shown in Figs. 8.3 and 8.4. Lysozyme molecular weight obtained from the intercept of the Zimm plot in the absence of PEG ($c_2 = 0$) is 14434 Da, and the slope gives $B_{11} = 27.8 \times 10^{-4}$ mol ml/g$^2$. This osmotic second virial coefficient differs by a factor of about two compared to the osmotic second virial coefficient derived from our previous light scattering study of aqueous lysozyme solutions at low ionic strength and the same pH and bis-Tris buffer: $B_{11} = 58.5 \times 10^{-4}$ mol ml/g$^2$ [23]. The difference is attributed to the different instruments used in the two studies.

The hPEG and mPEG concentration dependences of $M_{1}^{app}$ and $B_{11}^{app}$ are shown in Figs. 8.5 and 8.6 respectively. Osmotic cross virial coefficients derived from these results are compiled in Table 8.1. The hPEG-lysozyme cross virial coefficients at high ionic strength are compared in that table with results obtained previously in a similar light scattering study [236]. The $B_{12}$ obtained here is lower than the $B_{12}$ reported
Table 8.1: The osmotic virial coefficient, $B_{11}$, osmotic cross virial coefficients, $B_{12}$ and $C_{112}$, and the polymer concentration dependence of the effective osmotic virial coefficient, $B_{11}^{\text{eff}}$, for aqueous solutions of lysozyme (1) and hydroxy-terminated PEG (hPEG, 2) or methyl-terminated PEG (mPEG, 2) at ionic strengths of 0.01 and 0.8 M and $T = 25^\circ\text{C}$.

<table>
<thead>
<tr>
<th>polymer terminal group</th>
<th>ionic strength</th>
<th>$B_{11} \times 10^4$ mol ml/g$^2$</th>
<th>$B_{12} \times 10^4$ mol ml/g$^2$</th>
<th>$C_{112} \times 10^4$ mol ml$^2$/g$^3$</th>
<th>$\partial B_{11}^{\text{eff}}/\partial c_2 \times 10^4$ mol ml$^2$/g$^3$</th>
<th>ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>hPEG-lysozyme solutions</td>
<td>0.01</td>
<td>27.8</td>
<td>$\sim 0$</td>
<td>30.3</td>
<td>91</td>
<td>this work</td>
</tr>
<tr>
<td>hPEG</td>
<td>0.01</td>
<td>58.5</td>
<td>$\sim 0$</td>
<td>$\sim 0$</td>
<td>$\sim 0$</td>
<td>$[23]$</td>
</tr>
<tr>
<td>hPEG</td>
<td>0.8</td>
<td>-7.85</td>
<td>3.4</td>
<td>13.6</td>
<td>38</td>
<td>this work</td>
</tr>
<tr>
<td>hPEG</td>
<td>0.8</td>
<td>-4.85</td>
<td>12.9</td>
<td>16.1</td>
<td>16</td>
<td>$[236]$</td>
</tr>
<tr>
<td>mPEG-lysozyme solutions</td>
<td>0.01</td>
<td>27.8</td>
<td>-10.3</td>
<td>134.9</td>
<td>384</td>
<td>this work</td>
</tr>
<tr>
<td>mPEG</td>
<td>0.8</td>
<td>-7.85</td>
<td>$\sim 0$</td>
<td>15.2</td>
<td>46</td>
<td>this work</td>
</tr>
</tbody>
</table>

$^a$interpolated from data at ionic strengths of 0.007 M and 0.05 M.
previously, while $C_{112}$ obtained here is in good agreement with the previously reported value. In addition, $\partial B_{11}^{\text{eff}}/\partial c_2 > 0$, which was likewise reported in [236], indicating that the addition of hPEG makes a net repulsive contribution to the lysozyme potential of mean force.

From Fig. 8.5, the apparent molecular weight of lysozyme is only weakly dependent on PEG concentration at both ionic strengths for hPEG and at high ionic strength for mPEG. The weak dependence is reflected in a small positive $B_{12}$ for hPEG-lysozyme at 0.8 M, and $B_{12} \sim 0$ for the other two solutions (Table 8.1), which implies that PEG-lysozyme potentials of mean force are insignificant for hPEG and for mPEG at high ionic strength. The one exception to this weak dependence is the striking mPEG dependence of the lysozyme apparent molecular weight at 0.01 M. In this case, $B_{12} = -10.3 \times 10^{-4}$ mol ml/g$^2$ indicates that mPEG-lysozyme attractive interactions become significant at 0.01 M.

The polymer concentration dependence of the apparent osmotic second virial coefficient in Fig. 8.6 is unambiguously positive for both polymers, although $B_{11}^{\text{app}}$ is more sensitive to polymer concentration at low ionic strength. Again, it is the strong concentration dependence for mPEG at 0.01 M that stands out. Consequently, $C_{112}$ for the mPEG-lysozyme solution at this ionic strength is roughly an order of magnitude greater than $C_{112}$ for both PEG-lysozyme solutions at 0.8 M, and more than four times greater than $C_{112}$ for the hPEG-lysozyme solution at 0.01 M (Table 8.1). The strong mPEG concentration dependence is likewise reflected in $\partial B_{11}^{\text{eff}}/\partial c_2$, which is dominated by the value of $C_{112}$ (Eq. 8.7). This partial derivative is similarly four times more positive for mPEG than it is for hPEG at the lower ionic strength.
The effect of PEG end-group hydrophobicity probed here by replacing the hydrophilic end-groups of hPEG with the hydrophobic end-groups of mPEG at fixed PEG molecular weight is not expected to significantly impact the magnitude of steric repulsion as a contribution to the lysozyme potential of mean force. Light scattering from the two aqueous PEG solutions detected no clustering of mPEG due to the formation of interchain hydrophobic bonds (data not shown), as has been observed experimentally for higher molecular-weight PEG [244]. Loop closure by intra-chain hydrophobic bonding of the mPEG end-groups is also expected to have a negligible impact [245]. The PEG concentrations considered here are also below the crossover concentration from dilute to semi-dilute PEG solution behavior [237]. The predominant effect of PEG end-group hydrophobicity observed here at low ionic strength is to enhance mPEG-lysozyme attractive interactions, as reflected in $B_{12} < 0$ for mPEG – presumably, through the formation of intermolecular hydrophobic bonds between the end-groups of mPEG and the protein. This attraction, in turn, enhances mPEG-lysozyme preferential interactions. The preferential exclusion of lysozyme, and consequently, a lysozyme potential of mean force that becomes increasingly more repulsive follows directly from enhanced PEG-lysozyme preferential interactions.

The addition of low molecular-weight PEG to aqueous lysozyme solutions weakens lysozyme-lysozyme attractions, in apparent contradiction to the depletion-attraction mechanism, as was pointed out previously [236]. This effect is independent of solution ionic strength and PEG end-group hydrophobicity. For either polymer, increasing ionic strength reduces the impact of the added PEG, as expected [233]. The effect
of increasing end-group hydrophobicity is found to be comparable in magnitude, although opposite in direction (Table 8.1), which suggests that these hydrophobic interactions can play an important role in modulating protein-protein potentials of mean force. Other, potentially interesting outcomes might be anticipated for different end-group functionalities with different binding affinities and/or selectivities for specific sites on protein surfaces. As a practical matter, mixtures of low molecular-weight PEGs with multiple end-group functionalities could be used to “fine-tune” protein-protein interactions as a means to control phase behavior. Manipulating hydrophobic interactions alone – e.g., by using bulky fluorinated end-groups [246, 202] – might be an interesting strategy to pursue for directing site-specific protein-PEG interactions.

8.5 Conclusions

The addition of low molecular-weight PEG stabilizes aqueous lysozyme solutions by enhancing lysozyme-lysozyme repulsive interactions, independent of the solution ionic strength and PEG end-group hydrophobicity. The impact of adding mPEG vs. hPEG on stabilization is seen at low ionic strength, where the hydrophobicity of the mPEG end-groups enhances mPEG-lysozyme attractive interactions. These attractions promote preferential interactions of PEG with lysozyme at the expense of lysozyme-lysozyme interactions, which makes the lysozyme potential of mean force more repulsive. The findings highlight important role PEG-protein attractive interactions can play in mediating protein-protein interactions in aqueous solution.
CHAPTER 9

Static Light Scattering Measurements of Protein Osmotic Cross Virial Coefficients

9.1 Introduction

With the increasing use of protein therapeutics, accurate, quantitative knowledge of protein solution thermodynamics is essential for bulk production of proteins. One example is successive protein precipitation and crystallization used in pharmaceutical industry for the production of high purity proteineous drugs. Protein-protein interactions, characterized by osmotic virial coefficients, can be an important design parameter for protein phase separation. It has been observed that proteins crystallize when the solution osmotic second virial coefficient falls within the range of $-1 \times 10^{-4}$ to $-8 \times 10^{-4}$ mol ml/g$^2$ and precipitate when the osmotic second virial coefficient is below $-8 \times 10^{-4}$ mol ml/g$^2$. Since industrial protein separations involve fractionation of proteins from protein mixtures, it is crucial to understand the cross interactions between various proteins.

Osmotic second virial coefficients of a single protein solution can be measured using a wide variety of techniques: static light scattering, small angle neutron scattering, dynamic light scattering, small angle X-ray scattering, membrane osmometry, sedimentation
equilibrium \[258\ \text{and self-interaction chromatography} \ [260\ 261\ 262\ 263\ 264].

However, experimental techniques to measure osmotic second cross virial coefficients characterizing the cross interactions between different protein species in a mixture are not well developed.

Protein cross-association has been measured by membrane osmometry \[265\ 266\], static light scattering \[267\ 268\ 269\], sedimentation equilibrium \[270\ 271\ 272\], size exclusion chromatography \[273\ 274\ 275\ 276\], and fluorescence depolarization \[277\]. In these studies, the extent of complex formation and the corresponding association equilibrium constants are estimated. Association constants are well suited for strong protein-protein interactions described by a specific stoichiometry. Protein phase behavior on the other hand is determined by weak protein-protein interactions over an ensemble of protein configurations, best characterized by osmotic virial coefficients.

Cross interaction chromatography (CIC) is currently used to directly measure the osmotic second cross virial coefficients between different protein species \[278\ 279\ 280\]. Here, one protein is covalently immobilized on a chromatography column and the other is passed through the column in solution. The retention time of the protein in solution on the column is a measure of protein-protein interaction. This technique assumes that all possible configurations of the immobilized protein are sampled by the protein in solution. An accurate sampling of all protein configurations may not be guaranteed. It is also possible that non-specific interaction between the immobilized protein and the column packing material could modify the stable native configurations. The influence of the neighboring immobilized proteins cannot be neglected at a high surface coverage of proteins or at low salt concentrations or when the size difference between the immobilized protein and the protein in solution
are large. Also, the retention time depends on the concentration of the injected protein solution. In order to eliminate the concentration dependence, CIC requires the injection of high protein concentrations. Preparation and maintenance of the chromatography column are also tedious.

It is believed that static light scattering can not be used to measure the second cross virial coefficients directly [278, 279]. The statistical thermodynamic framework for the analysis of binary polymer solutions in terms of cross virial coefficients was, however, established in early 50’s [242, 241]. Recently, Bloustine et al., [236] applied the theory to light scattering measurements of the cross interactions between lysozyme and low molecular weight PEG 1000 and 8000 Da, to test depletion-attraction theory [281, 282, 283]. In this study, we extend their analysis to measure the osmotic second cross virial coefficient between lysozyme and α-chymotrypsinogen A.

9.2 Theory

Statistical mechanical analyses of light scattering from multicomponent solutions were derived by Brinkman and Hermans [284], Kirkwood and Goldberg [242] and Stockmayer [241]. The turbidity due to the Rayleigh scattering (inelastic scattering) of light from multicomponent solution is given [241] by,

\[
\tau = \frac{32\pi^2 n^2 k_B T}{3\lambda^4} V \sum_{i \geq 2} \sum_{j \geq 2} \psi_i \psi_j \left( \frac{\partial m_i}{\partial \mu_j} \right)_{T,p,m_1,\mu} ,
\]

where \(m_i\) and \(\mu_i\) are the number of molecules and the chemical potential of component \(i\), respectively, \(n\) is the refractive index of the solution, \(\lambda\) is the wavelength of the light, \(k_B T\) is the thermal energy, \(V\) is the volume, and \(\psi_i = \partial n / \partial m_i\) is the refractive index increment of component \(i\) at constant temperature and pressure. The summation excludes the pure solvent (component 1). The subscripts, \(\mu\) and \(m_1\) on the
partial derivative specify that this partial derivative is evaluated at constant chemical potentials of all components, except the solvent, for which the number of molecules, is held fixed. It is convenient to determine \( \partial \beta \mu_i / \partial \rho_j \) instead of \( \partial m_j / \partial \mu_i \), hence, Eq. 9.1 is rewritten as,

\[
\tau = \frac{32\pi^3 n^2}{3\lambda^4 N_A} \sum_{i \geq 1} \sum_{j \geq 1} \Psi_i \Psi_j \frac{A_{ij}}{|a_{ij}|},
\]

where \( \rho_i \) is the molar density of component \( i \), \( \beta = 1/k_BT \), \( \Psi_i \) is the derivative of the refractive index with respect to molar density of component \( i \), \( N_A \) is Avogadro’s number, \( |a_{ij}| \) is the determinant of the coefficients, \( a_{ij} \) defined by

\[
a_{ij} = \left( \frac{\partial \beta \mu_i}{\partial \rho_j} \right)_{T,p,\rho_k} = \left( \frac{\partial \beta \mu_j}{\partial \rho_i} \right)_{T,p,\rho_k},
\]

and \( A_{ij} \) is the cofactor of the element \( a_{ij} \).

For a single protein in solution, light scattering from the much larger protein molecules dominate over the scattering from the solvent. For this solution, the chemical potential [285] of the protein (component 2) is expressed as,

\[
\beta \mu_2 = \beta \mu_2^0 + \ln \rho_2 + 2B_{22}\rho_2 + 3C_{222}\rho_2^2 + \ldots,
\]

where \( B_{22} \) and \( C_{222} \) are the second and third virial coefficients, corresponding, respectively, to the pair and triplet potential of mean force for the protein molecules. Retaining only the second virial coefficient and substituting into Eq. 9.2 gives the well-known equation for a Zimm analysis:

\[
\frac{Kc_2}{R_{90}} = \frac{1}{M_2} + 2B_{22}c_2,
\]

where \( K \) is equal to \( 4\pi^2 n^2 (\partial n/\partial c_2)^2 /\lambda^4 N_A \), \( c_2 \) is the concentration of the protein in mg/ml, and \( R_\theta \) is the Rayleigh ratio which is related to turbidity: \( R_\theta = 6\tau / 16\pi (1 + \cos^2 \theta) \). At a scattering angle of 90°, \( R_{90} = 6\tau / 16\pi \).
For a solution containing two proteins, light scattering from both proteins must be considered and $\Psi_2 \sim \Psi_3$. It is convenient to analyze the turbidity of this solution ($\tau_{2+3}$) in excess of the turbidity of a solution containing either proteins, e.g. ($\tau_3$). The excess turbidity is given by

$$\Delta \tau = \tau_{2+3} - \tau_3 = \frac{32\pi^3 n^2 / 3 \lambda^4 N_A}{\partial \beta \mu / \partial \rho} \frac{1 - 2\alpha \omega_1 + \alpha \omega_1^2}{1 - \omega_1 \omega_2}, \quad (9.6)$$

where,

$$\omega_1 = (\partial \beta \mu_3 / \partial \rho_2) / (\partial \beta \mu_3 / \partial \rho_3);$$

$$\omega_1 = (\partial \beta \mu_3 / \partial \rho_2) / (\partial \beta \mu_2 / \partial \rho_2);$$

$$\alpha = \Psi_3 / \Psi_2.$$

The chemical potential of one protein in this solution depends on the concentration of both proteins, and the self and cross interactions: [236]

$$\beta \mu_i = \beta \mu_i^0 + \ln \rho_2 + 2B_{ii} \rho_i + 2B_{ij} \rho_j + 3C_{iii} \rho_i^2$$

$$+ 6C_{iij} \rho_i \rho_j + 3C_{ijj} \rho_j^2 + \ldots. \quad (9.7)$$

The coefficients in the series expansion include the self interaction coefficients – second virial coefficient, $B_{ii}$, and third virial coefficient, $C_{iii}$, – and the cross interaction coefficients – second cross virial coefficient, $B_{ij}$, and third cross virial coefficients, $C_{iij}$ and $C_{ijj}$.

The Zimm equation for this solution is:

$$\frac{Kc_2}{(R_{90}^{2+3} - R_{90}^2)} = \frac{Kc_2}{\Delta R_{90}} = \frac{1}{M_{app}^2} + 2B_{22}^{app} c_2, \quad (9.8)$$

where, $c_i$ is the concentration of the protein $i$ in mg/ml, $M_{app}^2$ is the apparent molecular weight of one protein (component 2), $B_{22}^{app}$ is the apparent second virial coefficient describing an effective interaction between two molecules of this protein in
the presence of the second protein (component 3). Eq. 9.8 is similar in the form to Eq. 9.5, except that the thermodynamic quantities obtained from the intercept and the slope defined by Eq. 9.8 are apparent quantities containing the information on cross interaction between species. Thus cross interactions between the proteins can be determined [242, 236] by measuring the difference between excess Rayleigh ratio of the binary protein solution, $R_{90}^{2+3}$, and the excess Rayleigh ratio of a single protein containing only one of the either proteins in the solution, $R_{90}^{3}$, as a function of $c_2$.

The apparent molecular weight and the apparent second virial coefficient depend on the concentration of the second protein,

$$\frac{1}{M_2^{\text{app}}} = \frac{1}{M_2} + 4mB_{23}c_3;$$

(9.9)

$$B_{22}^{\text{app}} = B_{22} + (3mc_{223} - 2mB_{23}^2M_3) c_3 + (6mc_{223} - 4mB_{22}B_{23}M_2) c_3,$$

(9.10)

where, $m = M_3(\partial n/\partial c_3)/M_2(\partial n/\partial c_2)$ with $n$ denoting the refractive index of the solution and $M_i$ to the molecular weight of the protein $i$. The cross virial coefficients can be obtained directly through the light scattering experiments from the measurement of apparent molecular weight of one protein in presence the second protein.

### 9.3 Materials and Methods

Hen egg white lysozyme (3× crystallized, Catalog No. L6876) and α-chymotrypsingen A from bovine pancreas (6× crystallized, Catalog No. C4879) were purchased from Sigma Aldrich and used without any further purification. Bis-Tris buffer (bis-Tris hydrochloride: B6032; bis-Tris: B9754, Sigma Aldrich) was prepared using the recipe from buffer calculator of University of Liverpool (http://www.liv.ac.uk/buffers/) to
maintain a protein solution pH of 7.0. The pH was measured using a OAKTON Acron Ion 6 pH meter and was adjusted to 7.0 by adding small volumes of 0.1 M HCl (H1758, Sigma Aldrich) or NaOH (480878, Sigma Aldrich). The ionic strength of the solution was adjusted to 0.1 M using NaCl (Z273287, Sigma Aldrich).

All solutions were prepared with filtered deionized water using a Barnstead Nanopure Ultraviolet water filter system (Barnsted International). Buffer solutions were filtered using Millipore vacuum membrane filtration unit containing 0.1 µm PVDA membrane prior to preparing protein stock solutions. All glassware was first treated with detergent, stored overnight in HELLMAMEX II alkaline cleaning solution, and then washed thoroughly with filtered, deionized water shortly before an experiment. To minimize protein aggregation, the protein solutions were prepared less than 2 hrs before the measurements and filtered using Amicon Ultrafree MC centrifugal filter devices (Millipore, Billerica, MA) with a pore size of 0.1 µm to remove any particulates and aggregates.

Scattered light was collected at an angle of 90° using miniDawn Treos (Wyatt Technology, CA) equipped with Gallium Arsenide (GaAs) laser operating at a wavelength of 658 nm. The refractive index increment of both proteins – ∂n/∂c2 – was taken to be 0.185 ml/g. A composition gradient pump system (Calypso 1.2.3.6), equipped with three 1 ml syringe-piston pumps, controls the protein composition during an experiment by modifying the injection volumes of the protein solution and a buffer solution. The concentration of protein(s) in the solution was determined by UV absorption at 280 nm using a Varion Prostar 325 UV spectrophotometer connected in series up stream of the static light scattering detector. The UV absorption extinction coefficients at 280 nm were taken to be 2.64 liter/g cm and 2.0 liter/g
Figure 9.1: Zimm plots (Eq. [9.8]) for aqueous solutions of lysozyme (2) and α-chymotrypsinogen A (3). Symbols correspond to the experimental data and the lines are linear fits, at each constant concentration of α-chymotrypsinogen A. Symbols represent different concentrations of α-chymotrypsinogen A: solid circles - 0 mg/ml; open circles - 0.45 mg/ml; solid squares - 0.89 mg/ml; open squares - 1.64 mg/ml; solid triangles - 2.57 mg/ml; open triangles - 3.49 mg/ml.

cm for lysozyme and α-chymotrypsinogen, respectively. All experiments were carried out at 25°C. After equilibration with buffer for about 2 hours, as determined by a noiseless baseline, the run-time for each experiment was about 2 hours.

9.4 Results and Discussion

Osmotic second cross virial coefficients for lysozyme and α-chymotrypsinogen A, were obtained at a solution pH of 7.0 and an ionic strength of 0.1 M (including buffer ions along with the ions of the salt NaCl) from the experimental data shown in Figs. [9.1] and [9.2]. We measured the difference between the excess Rayleigh ratio of a
Figure 9.2: Zimm plots (Eq. 9.8) for aqueous solutions of lysozyme (2) and \( \alpha \)-chymotrypsinogen A (3). Symbols correspond to the experimental data and the lines are the linear fits, at each constant concentration of lysozyme. Symbols represent different concentrations of lysozyme: solid circles - 0 mg/ml; open circles - 0.52 mg/ml; solid squares - 0.88 mg/ml; open squares - 1.68 mg/ml; solid triangles - 2.83 mg/ml; open triangles - 3.58 mg/ml.
binary solution of lysozyme and α-chymotrypsinogen A and the excess Rayleigh ratio of solution containing either α-chymotrypsinogen A (Fig. 9.1) or lysozyme (Fig. 9.2) at a specified concentration. With the increase in the concentration of the second protein, the intercept decreases in both Figs. 9.1 and 9.2. However, the trend is different for the slope of Zimm plots. The slope of the binary solution, as shown in Fig. 9.1, is strikingly opposite to that of the solution containing only lysozyme. Whereas, the slope is negative up to an addition of 1 mg/ml lysozyme in the solution of α-chymotrypsinogen A, while it becomes positive beyond 1 mg/ml (Fig. 9.2).

The intercepts of Zimm plots (Eq. 9.5) for single protein solutions (data shown as solid circles in Figs. 9.1 and 9.2) correspond to molecular weights of 14716 Da for lysozyme and 24682 Da for α-chymotrypsinogen A, in good agreement with the molecular weights of 14300 Da and 25600 Da, respectively, from the amino acid
sequence. The second virial coefficients for the individual proteins in solution were determined to be \(-4.1 \times 10^{-4}\) and \(-4.6 \times 10^{-4}\) for lysozyme and \(\alpha\)-chymotrypsinogen A, respectively. The negative values imply attractive self interactions. Velev et al., [15] measured virial coefficients for lysozyme and \(\alpha\)-chymotrypsinogen A at a solution pH of \(\sim 7\) and 0.1 M NaCl using light scattering: \(-3.65 \times 10^{-4}\) mol ml/g\(^2\) (pH of 7.5) and \(-4.1 \times 10^{-4}\) mol ml/g\(^2\) (pH of 6.8), respectively, are in good agreement with the measured values obtained in this work.

The linear fits to the data for binary solutions in Figs. 9.1 and 9.2 provide the inverse of apparent molecular weight as the intercept and the twice the apparent second virial coefficient as the slope (Eq. 9.8). Fig. 9.3 shows the change in the apparent molecular weight of lysozyme when \(\alpha\)-chymotrypsinogen A is added into the solution and vice-versa. The intercepts and slopes of the linear fits in that figure provide molecular weight of the protein and the second cross virial coefficient characterizing the cross interaction between the proteins, as given by Eq. 9.9.

The molecular weights obtained from this analysis – 19167 Da for lysozyme and 34372 Da for \(\alpha\)-chymotrypsinogen A – are over-estimated by about 30% relative to those calculated based on amino acid sequences. The higher molecular weights presumably reflect a significant difference in the local environment of either protein as the second protein becomes infinitely dilute compared to the local solution environment of the pure protein itself.

The osmotic cross virial coefficients for lysozyme-\(\alpha\)-chymotrypsinogen A (Table. 6) obtained in the two experiments are comparable, and to the literature value measured by cross interaction chromatography [278]. The statistical error in the
Table 9.1: Comparison of osmotic virial coefficients corresponding to self and cross interactions between lysozyme, α-chymotrypsinogen A measured by static light scattering (SLS) to the literature values.

<table>
<thead>
<tr>
<th>Protein</th>
<th>$B_{22} \times 10^4$</th>
<th>ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lysozyme</td>
<td>-4.1</td>
<td>this work</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>-3.6$^a$</td>
<td>[15]</td>
</tr>
<tr>
<td>α-chymotrypsinogen A</td>
<td>-4.6</td>
<td>this work</td>
</tr>
<tr>
<td>α-chymotrypsinogen A</td>
<td>-4.1$^b$</td>
<td>[15]</td>
</tr>
<tr>
<td>Proteins</td>
<td>$B_{23} \times 10^4$</td>
<td>ref</td>
</tr>
<tr>
<td>Lysozyme/α-chymotrypsinogen A</td>
<td>-11.3</td>
<td>this work</td>
</tr>
<tr>
<td>Lysozyme/α-chymotrypsinogen A</td>
<td>-11.8$^c$</td>
<td>[278]</td>
</tr>
<tr>
<td>α-chymotrypsinogen A/Lysozyme</td>
<td>-16.1</td>
<td>this work</td>
</tr>
</tbody>
</table>

$^a$static light scattering data for lysozyme at 7.5 pH and 0.1 M ionic strength.

$^b$static light scattering data for α-chymotrypsinogen A at 6.8 pH and 0.1 M ionic strength.

$^c$cross interaction chromatography data when 25 mg/ml lysozyme was injected into a column with immobilized α-chymotrypsinogen A at 7 pH and 0.15 M ionic strength.
The measured osmotic second cross virial coefficient is $\sim \pm 3 \times 10^{-4} \text{ mol ml/g}^2$. The negative cross virial coefficient reveals an attractive interaction between lysozyme and $\alpha$-chymotrypsinogen A, despite the fact that both the proteins have a net positive charge at pH 7.

More accurate determinations of the molecular weights and osmotic virial coefficients require measurements at very low protein concentration ($< 2 \text{ mg/ml}$), as the contribution of proteins to the total ionic strength of the solution cannot be neglected at high concentrations [23], which would lead to non-linear Zimm plots. We have measured the osmotic second cross virial coefficients for a protein concentration range of 0.5–4 mg/ml. Extension of the measurements to lower concentrations would provide insights on the cause for the over-estimation of the protein molecular weights obtained from the linear fits in Fig. 9.3. The measurement at low concentrations was, however, limited by the uncertainty in the accurate automated dilution using composition-gradient system (containing 1 ml injection syringes) as complete mixing may not be ensured for large differences in injection volumes for protein and buffer solutions, also, at low protein injection volumes the protein loss due to adsorption on the mechanical parts of the injection port cannot be neglected. Nevertheless, the highest concentration (4 mg/ml) measurement in this work is about six times lower than the lowest concentration (25 mg/ml) measurement possible by CIC measurement.

CIC measurements assume that immobilized proteins exist in all solution configurations, at least 50% of immobilized protein surface is accessible for the mobile proteins, and that one immobilized protein molecule interacts with one mobile protein molecule without influence from neighboring protein molecules. In practice, however, the immobilized protein may adopt different configurations upon adsorption on a
solid surface. Also, the CIC measurement may not sample important configurations of either protein as required for the osmotic virial coefficient determination. Measurements at low ionic strength are also restricted due to interfering long-range interactions with neighboring immobilized protein molecules. Static light scattering does not require these assumptions, and is a direct measure of solution property of interest. In addition CIC measurements are based on a single protein concentration, with high possibility of concentration dependent measurements [278, 279]. In contrast, light scattering measures $B_{23}$ as an average property over a range of concentration. Hence, the time and amount of proteins required for these measurements are larger than required for CIC.

9.5 Conclusions

To date, cross interaction chromatography is considered to be the only experimental technique for direct measurements of osmotic cross virial coefficients. Though static light scattering technique have been used extensively to measure osmotic virial coefficients for a single protein in solution, its application to measure osmotic cross virial coefficients have only recently been explored.

In this work, we have shown the direct measurement of osmotic second cross virial coefficients between lysozyme and $\alpha$-chymotrypsinogen A by static light scattering. The measured cross virial coefficient between lysozyme and $\alpha$-chymotrypsinogen A compare reasonably well to CIC measurements. At the solution condition of pH 7 and 0.1 M ionic strength, both the self interactions and cross interaction between the proteins are attractive, however, the attraction between lysozyme and $\alpha$-chymotrypsinogen A is about thrice that of their corresponding self attractions.
The molecular weights of the proteins evaluated from the Zimm analysis for binary solutions, was over-estimated by 30%, presumably reflecting the difference in the local environment of either proteins in binary solutions compared to the local environment of the pure protein itself.
CHAPTER 10

Conclusions and Future work

10.1 Conclusions

The principal objective of the work presented in this thesis is to understand the fundamental molecular level interactions among protein, water and co-solvents – specifically, alcohols and polyols – in exploring the role of hydration and co-solvent preferential interactions on protein-protein interactions. Quasi chemical theory [29, 213, 214] distinguishes the local interactions of the solute with the water and co-solvent from the non-local interactions, partitioning the solvent medium into two parts: i) strongly interacting close neighbors identified by occupancy of water and co-solvent molecules within a defined proximal volume of inner shell around a protein site, and ii) a bulk medium or an outer shell comprising the rest of the solvent molecules. The theory, therefore, provides an appropriate molecular framework to investigate the local hydration and co-solvent preferential interactions at specific sites of topologically and chemically complex solutes such as globular proteins.

In the quasi-chemical view of protein hydration, it is natural to consider strongly associated water molecules as a part of the protein [29]. Specific hydration sites on the surface of globular proteins are defined in terms of the local water density at each site relative to bulk water density [23, 22]. Alternatively, the specific hydration sites can
also be characterized by the residence times of the water at the site [41]. The analysis of average water residence times and vacancy times at specific hydration sites on the protein surface revealed two distinct kinetic regimes for those hydration sites defined thermodynamically to have high local water densities: long residence times relative to vacancy times for a single water molecule, corresponding to kinetically bound water molecules, and short residence times with high turnover involving multiple water molecules. Those sites corresponding to kinetically bound water molecules comprise only a small fraction of the total number of high occupancy sites, and are correlated with local heterogeneities in both surface charge and roughness. Moreover, these sites have little impact on calculated osmotic second virial coefficients for protein-protein interactions. The impact of preferential hydration on these weak protein-protein interactions is due primarily to the preferential hydration of sites characterized by high occupancy and high turnover – i.e., those sites on the protein surface that are accessible to water.

The effect of surface curvature on protein hydration is a least understood phenomenon. We investigated the curvature effect on hydration through a model system of a large sinusoidal hydrophobic wall. When the interaction between wall and water is attractive, the surface remains hydrated. With decrease in wall-water attractions the surface dewets; the effect was profound in concave wells than on convex crest. Our result reinforces the presumption of stronger cavity-expulsion potential on a concave surfaces [80]. Thus the hydration of hydrophobic patches on protein surfaces, usually found in the concave pockets, will be strongly determined by its local chemical composition.
The protein atoms have been chemically classified into a minimum set of 11 molecular groups, and local water structure of water around amino acids and specific protein sites have been obtained from the proximal distribution of water around the constituent molecular groups. Partial molar volume of amino acid side chains computed relative to glycine from the local water structure using Kirkwood-Buff theory closely matches the experimental data. The proximal distribution of water around identical molecular groups across polar/non-polar amino acids are indistinguishable, however, the proximal distribution in case of charged amino acids are distinctly different, due to long range electrostatic interactions.

The molecular preferential interactions of organic co-solvents – methanol, ethanol, urea and glycerol – with molecular solutes such as methane, ethane, neopentane and tetra-methyl ammonium ion, and specific sites on lysozyme have been quantified in terms of preferential interaction parameter using Kirkwood-Buff and quasi-chemical theories. Kirkwood-Buff theory determines the preferential interaction parameter on the basis of the difference in the molecular distributions of co-solvent and water molecules in the vicinity of the solute relative to their bulk distribution. The characteristic difficulty in the reliable estimate of preferential interaction parameter that arises from long range oscillations in the radial distribution functions (rdfs), derived from molecular simulations, have been surpassed by smearing the distribution functions assuming multiple solvent reference centers around each heavy atom of the solvent molecules. The preferential interaction parameters evaluated as the difference of smeared rdfs converge about 10 – 12 Å from the solute center. The local preferential interaction parameters at specific sites of a solute have also been computed by smearing the distribution functions of co-solvent and water proximal to the sites.
Quasi chemical theory provides an alternate approach in evaluating the co-solvent preferential interaction of solute in aqueous solutions. The advantage of the theory is that the calculated preferential interaction parameter is independent of the inner shell volume around the solute and can be estimated accurately using inner shell volumes that are significantly smaller than those required through Kirkwood-Buff theory. The inner-shell local interactions are described by the chemical association of a solute molecule with molecular clusters of the solvent averaged over a statistical distribution of solvent compositions. Solute interactions with outer-shell solvent molecules make an important contribution by preferentially stabilizing certain inner shell solvent clusters over others.

The partitioning of local interactions and non-local interaction in QC theory naturally enables estimation of co-solvent preferential interactions at specific sites on protein surface, which is of primary importance to protein-protein interactions in aqueous solutions. The local preferential interactions at a protein site is more informative than a preferential interaction parameter averaged over the entire protein surface. Although the local preferential interaction parameters can be estimated from the proximal distribution of co-solvent and water for the site, the technique neglects the long-range nature of the electrostatic interactions, hence, could not provide a detailed picture on preferential interactions. Considering the polar nature of investigated protein sites, it is necessary to account for the electrostatic interactions; quasi chemical theory clearly reveals the importance of these interactions as they contribute dominantly to the local preferential interactions of the protein sites.
The effects of PEG end-group hydrophilicity/hydrophobicity on lysozyme-lysozyme interactions in aqueous solution of hydroxy-terminated PEG-1000 (hPEG) and methyl-terminated PEG-1000 (mPEG) were investigated at two solution conditions – pH 7 and 0.01 M ionic strength, and pH 6.2 and 0.8 M ionic strength – through static light scattering. The addition of polymer stabilizes the dissolution of lysozyme by enhancing the lysozyme-lysozyme repulsive interactions, independent of end-group functionality and solution ionic strength. The effect of PEG end group hydrophobicity is profound at low ionic strength where mPEG-lysozyme attractive interactions become significant, and the net contribution to the lysozyme potential of mean force is much greater than that for hPEG. Thus, the predominant effect of increasing the end-group hydrophobicity of PEG is to promote preferential interaction of PEG with lysozyme, which makes a net repulsive contribution to the lysozyme potential of mean force.

Knowledge of protein-protein interactions between different proteins in solution is crucial for the design of successful protein fractionation processes. Weak interaction between different proteins are characterized by osmotic cross virial coefficients. To date, cross interaction chromatography is the only experimental technique for measuring osmotic second cross virial coefficients. We measured the osmotic second cross virial coefficient between lysozyme and \( \alpha \)-chymotrypsinogen A by static light scattering based on the statistical thermodynamic theory of multi-component solutions. Static light scattering enables measurements at low protein concentrations and at low ionic strength without any assumptions about the sampling all possible protein configurations.
The overall conclusions reached from this research are the following: First, the surface roughness and curvature play a dominant role in protein hydration, in addition to the widely studied chemical heterogeneity. Also, the dynamics of water in high occupancy sites has little impact on protein-protein interactions. Second, quasi-chemical theory which distinguishes local and non-local interactions naturally provides the appropriate framework for modeling local preferential hydration and co-solvent preferential interactions at specific sites on protein surface. Though, work required for cavity formation contributes dominantly for the preferential interaction of non-polar solutes, the outer shell interactions, especially the electrostatic interactions, contribute dominantly for polar or charged protein sites. Finally, fine tuning of hydrophobicity/hydrophilicity of PEG end-group to enhance/lower the preferential interaction with a protein, can stabilize/de-stabilize the dissolution of the protein in aqueous solutions.

10.2 Future work

Weak protein-protein interactions are well characterized by osmotic virial coefficients. As discussed in the chapter 1, osmotic second virial coefficient of protein solution determines the protein phase separation, hence, modeling the osmotic second virial coefficient will help in a rapid screening of protein phase behavior. The straightforward application of the molecular theories on co-solvent preferential interactions developed in this thesis is to incorporate it into the model for potential of mean force between proteins [23] to predict the osmotic second virial coefficient of protein at a given solution condition as discussed below.
According to the theory of light scattering from multicomponent solutions [284, 241, 242], the solution turbidity arising from the concentration fluctuations is given by

$$\tau = H \sum_{i \geq 1} \sum_{j \geq 1} \psi_i \psi_j \frac{A_{ij}}{|a_{ij}|}$$  \hspace{1cm} (10.1)

where $\psi_i = \frac{\partial n}{\partial \rho_i}$ is the refractive index increment upon adding component $i$ at constant temperature and pressure, $\rho_i$ is the molar density of component $i$, $H$ is an optical constant equal to $32\pi^3 n^2 / 3\lambda^4 N_A$, $N_A$ is Avogadro’s number, $|a_{ij}|$ is the determinant of the coefficients of $a_{ij}$ given by

$$a_{ij} = \left( \frac{\partial \beta \mu_i}{\partial \rho_j} \right)_{T,P,\rho_k} = \left( \frac{\partial \beta \mu_j}{\partial \rho_i} \right)_{T,P,\rho_k},$$  \hspace{1cm} (10.2)

and $A_{ij}$ is the cofactor of the element $a_{ij}$.

For example in case of a binary solution comprising a solute (subscript $s$) and solvent, say water (subscript $w$),

$$\frac{H \psi_s^2 \rho_s}{\tau} = \rho_s \left( \frac{\partial \beta \mu_s}{\partial \rho_s} \right)_{T,P} = 1 + 2B_s \rho_s.$$  \hspace{1cm} (10.3)

The second equality is obtained from the definition of chemical potential for the solute, at its infinite dilution,

$$\beta \mu_s = \beta \mu_s^0 + \ln \rho_s + 2B_s \rho_s + \ldots,$$  \hspace{1cm} (10.4)

where $B_s$ is the osmotic second virial coefficient of the solute.

Assuming protein solution as a three component solution: water including buffer ions (subscript $w$), protein (subscript $p$), and co-solvent (subscript $c$), and protein as the dominant light scatter, the expression for turbidity becomes

$$\frac{H \psi_p^2 \rho_p}{\tau} = \rho_p \left( \frac{\partial \beta \mu_p}{\partial \rho_p} \right)_{T,P,\rho_c} - \rho_p \left[ \frac{\left( \frac{\partial \beta \mu_p}{\partial \rho_c} \right)_{T,P,\rho_p}^2}{\left( \frac{\partial \beta \mu_c}{\partial \rho_c} \right)_{T,P,\rho_p}} \right].$$  \hspace{1cm} (10.5)
Recollecting the definition of preferential interaction parameter

\[
\Gamma_{pc} \equiv -\left( \frac{\partial \beta_{\mu_p}}{\partial \beta_{\mu_c}} \right)_{T,P,\rho_p} = -\left( \frac{\partial \beta_{\mu_p}}{\partial \rho_c} \right)_{T,P,\rho_p}, \quad (10.6)
\]

Eq. 10.5 simplifies to the form

\[
\frac{H\psi_p^2 \rho_p}{\tau} = \rho_p \left( \frac{\partial \beta_{\mu_p}}{\partial \rho_p} \right)_{T,P,\rho_c} - \rho_p \Gamma_{pc}^2 \left( \frac{\partial \beta_{\mu_c}}{\partial \rho_c} \right)_{T,P,\rho_p} = 1 + 2B_{p\rho_p}^\text{eff} \rho_p. \quad (10.7)
\]

The effective osmotic second virial coefficient \(B_{p\rho_p}^\text{eff}\) for the protein can be expressed as,

\[
1 + 2B_{p\rho_p}^\text{eff} = \rho_p \left( \frac{\partial \beta_{\mu_p}}{\partial \rho_p} \right)_{T,P,\rho_c} - \frac{\rho_p}{\rho_c} \Gamma_{pc}^2, \quad (10.8)
\]

assuming \(\partial \beta_{\mu_c}/\partial \rho_c \propto \rho_c^{-1}\).

Thus, by evaluating the preferential interaction parameter, \(\Gamma_{pc}\), either through quasi chemical theory or Kirkwood-Buff theory, the effective osmotic second virial coefficient of proteins can be evaluated. The chemical potential derivative in the first term on right of Eq. 10.8 quantifies the direct interaction between two protein molecules in the solution, can be calculated computationally for different orientations of the two protein molecules \[23, 22\]. The weak protein-protein interactions can also be investigated using coarse grain model for proteins \(i.e.,\) treating the amino acids as beads connected by elastic springs \[286\]. Such coarse grain models, which have been reported to capture the physics of protein binding occurring over longer time scales and even at larger length scales, will be helpful in investigating co-solvent induced protein denaturation.
Recent studies have reported converging KB integrals in estimating the partial molar volume of various solutes in water, by dampening the long range oscillations in the distribution of water around the solute, using a spherically smeared solvent distribution function \[34, 35, 134\]. The water solvent centers were assumed to be distributed uniformly within a spherical volume of radius 2.3 Å centered on water oxygen; the resulting KB integrals was the least oscillatory and converged at a distance of \(\sim 10\) Å from the solute center. For larger smearing volume, the convergence was delayed because of the interference with the periodic images of the simulation cell \[134\]. It is to be noted that 2.3 Å is the effective hard sphere diameter of the water oxygen \(i.e.,\) the distance up to which the neat water oxygen-oxygen radial distribution function is absolutely zero. Similarly, we found that when the distribution of methanol around the central methane molecule is smeared over a spherical radius of 3.3 Å about center of mass (COM) of methanol molecules, a converging estimate of partial molar volume of methane in methanol, within a few hydration shells, is obtained (Fig. \[A.1\]). The radial distance of 3.3 Å is the effective hard sphere diameter
of methanol COM such that in a neat methanol system, no two methanol COM-s can be seen within this distance.

The converging KB integrals of solvent distribution around several solutes - methane, ethane, neopentane and TMA - in neat water (left panel) and in neat methanol (right panel), computed using the smeared distribution of the solvent are shown in Fig. A.1. The partial molar volumes of the solutes calculated from the average of the integral values at a distance of 10 – 15 Å from the solute center, are reported in Table A.1 and compared against the experimental data. The uncertainty in the converging estimate of partial molar volume of the solutes are larger in methanol than that in water owing to the larger size of methanol and its orientational preference (methyl or hydroxyl group) towards the solute.
Figure A.1: KB integrals of the solvent around various solutes in neat water (top panel) and in neat methanol (bottom panel), as a function of cutoff distance, $R_c$, from the solutes, determined using the smeared effective grand canonical distribution of solvent around the solutes: methane - solid line; ethane - dashed line; neopentane - dash and dotted line; TMA - dotted line. The radii of 2.3 Å and 3.3 Å are used for smearing the distributions of water and methanol, respectively.
Table A.1: Partial molar volumes of molecular solutes(s) in neat water(i) and neat methanol(i) determined by KB approach: $\bar{V}_s^\infty = kT\kappa_0 - G_{si}$. The calculated values are compared against the values empirically predicted via group contribution technique based on the experimental data for a range of hydrophobic solutes.

<table>
<thead>
<tr>
<th>Solute</th>
<th>Solvent: Water $^a$</th>
<th>Solvent: Methanol $^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\bar{V}_s^\infty$</td>
<td>empirical group</td>
</tr>
<tr>
<td></td>
<td>calculated</td>
<td>contribution</td>
</tr>
<tr>
<td></td>
<td>$\bar{V}_s^\infty$</td>
<td>calculated</td>
</tr>
<tr>
<td></td>
<td>empirical group</td>
<td>contribution</td>
</tr>
<tr>
<td>Methane</td>
<td>38.8 ± 1.2</td>
<td>37.3 $^c$</td>
</tr>
<tr>
<td>Ethane</td>
<td>60.3 ± 0.7</td>
<td>51.5</td>
</tr>
<tr>
<td>Neopentane</td>
<td>105.9 ± 0.8</td>
<td>99.2</td>
</tr>
<tr>
<td>TMA</td>
<td>83.1 ± 0.5</td>
<td>–</td>
</tr>
</tbody>
</table>

$^a\kappa_0$ of water is $4.1 \times 10^{-4}$ MPa$^{-1}$ [133]

$^b\kappa_0$ of methanol is $12.1 \times 10^{-4}$ MPa$^{-1}$ [287]

$^c$values obtained from Ref. [288]

$^d$values obtained from Ref. [182]

$^e$experimental partial molar volume for methane in water

$^f$partial molar volume of methyl group in methanol
Figure A.2: Effect of concentration of methanol on the preferential interaction of methanol with neopentane (solid circles) and TMA (open circles). The lines are the linear fit to the data with $R^2 = 0.98$ for neopentane and $R^2 = 0.94$ for TMA.

Table A.2: Preferential interaction parameter of methanol with different solutes, in the units of number of cosolvent molecules per molecule of the solute, in 50 mol % methanol solution.

<table>
<thead>
<tr>
<th>Solute</th>
<th>$\Gamma_{sc}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methane</td>
<td>1.94 ± 0.72</td>
</tr>
<tr>
<td>Ethane</td>
<td>3.78 ± 0.48</td>
</tr>
<tr>
<td>Neopentane</td>
<td>4.58 ± 0.42</td>
</tr>
<tr>
<td>TMA</td>
<td>-5.39 ± 0.53</td>
</tr>
</tbody>
</table>
Figure A.3: Preferential interaction parameter of methanol with various solutes as a function of the cutoff distance, $R_c$, from the solutes in 50 mol % methanol solution calculated using the smeared effective grand canonical distributions of methanol and water. The different solutes are: methane - solid line; ethane - dotted line; neopentane - dashed line; TMA - dash-dotted line.
APPENDIX B

Supplementary Data for Quasi-Chemical Theory of Preferential Interactions in Aqueous Solutions

B.1 Error estimation of gaussian approximation for the outer-shell contribution to the solute excess chemical potential

Table B.1: Absolute values of the errors in the estimation of outer contribution by an gaussian approximation for the distribution of binding energies of methane, at different concentrations of methanol in the solution, in an empty inner shell defined by $-\lambda_w = 3.3 \text{ Å}$; $\lambda_m = 3.6 \text{ Å}$; $\lambda_h = 3.9 \text{ Å}$.

| methanol mole fraction | $|\Delta \mu_1| \times 10^2$ | $|\Delta \mu_2| \times 10^2$ |
|------------------------|-----------------------------|-----------------------------|
| 0.00                   | 0.69                        | 1.20                        |
| 0.05                   | 0.65                        | 1.95                        |
| 0.10                   | 0.85                        | 2.04                        |
| 0.20                   | 0.76                        | 1.63                        |
| 0.30                   | 0.77                        | 0.09                        |

The outer-shell contribution to the free energy of the solvation of a solute is

$$\ln \langle e^{+\beta \epsilon} | m, n = 0 \rangle = \ln \int_{-\infty}^{\infty} P(\epsilon | m, n = 0) e^{\beta \epsilon} d\epsilon .$$  \hfill (B.1)$$

When the probability distribution of solute binding energies is accurately gaussian,

$$P(\epsilon | m, n = 0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(\epsilon - \langle \epsilon \rangle)^2}{2\sigma^2} \right],$$  \hfill (B.2)$$
the outer-shell contribution is given by the sum of the mean binding energy and fluctuations in binding energy; i.e.,

\[ \ln \langle e^{+\beta \epsilon} | m, n = 0 \rangle = \langle \epsilon \rangle + \beta \sigma^2 / 2. \]  

(B.3)

Evaluation of this outer-shell contribution directly from Eq. B.1 requires integrating the binding energy probability distribution over the range \((-\infty, \infty)\). However, the binding energies are sampled only over a finite range \((\epsilon_{\text{min}}, \epsilon_{\text{max}})\), that is accessible in a simulation. The error, \(\Delta \mu_1\), associated with evaluating the outer-shell contribution assuming a gaussian probability distribution over the range \((-\infty, \infty)\) relative to numerically evaluating the integral in Eq. B.1 over the sampled, finite range of binding energies is

\[ \Delta \mu_1 = \langle \epsilon \rangle + \beta \sigma^2 / 2 - kT \ln \int_{\epsilon_{\text{min}}}^{\epsilon_{\text{max}}} P(\epsilon | m, n = 0) e^{\beta \epsilon} d\epsilon. \]  

(B.4)

It is observed that the gaussian fit to the binding energy probability distribution is appropriate only over a finite range \((\tilde{\epsilon}_{\text{min}}, \tilde{\epsilon}_{\text{max}})\). The exponential weighting factor \(e^{\beta \epsilon}\) in Eq. B.1 weights the high energy tail of the distribution dominantly over the low energy tail. Thus, with negligible loss of accuracy, we can assume the observed binding energy distribution to be accurate in the limit \((-\infty, \tilde{\epsilon}_{\text{max}})\). The error associated in assuming a gaussian distribution over the range \((\tilde{\epsilon}_{\text{max}}, \infty)\), however, may not be negligible and is estimated to be [136],

\[ \Delta \mu_2 = kT \ln \left( \frac{1}{2} \text{erfc} \left( \frac{\beta \sigma^2 - (\tilde{\epsilon}_{\text{max}} - \langle \epsilon \rangle)}{\sqrt{2} \sigma^2} \right) \right), \]  

(B.5)

where \(\text{erfc}\) is the complimentary error function.

174
Figure B.1: Conditional probability distributions of binding energies of methane in water when no waters are within: 3.3 Å (solid circles), 3.4 Å (open circles), 3.5 Å (solid squares), 3.6 Å (open squares), 3.7 Å (solid triangles), 3.8 Å (open triangles) of the center of the methane molecule. The energy distributions are successively shifted upwards by 2 units for visual clarity. Lines are gaussian fits to the data.
Figure B.2: Conditional probability distributions of binding energies of methane in methanol when no methanol methyl group is within a distance of $\lambda_m$ and no methanol hydroxyl group is within $\lambda_h = \lambda_m + 0.3$ Å of the center of the methane molecule for $\lambda_m$: 3.6 Å (solid circles), 3.7 Å (open circles), 3.8 Å (solid squares), 3.9 Å (open squares), 4.0 Å (solid triangles), 4.1 Å (open triangles). The energy distributions are successively shifted upwards by 2 units for visual clarity. Lines are gaussian fit to the data.
Figure B.3: Conditional probability distributions of binding energies of methane in aqueous methanol solutions at different methanol concentration when no water oxygens are within 3.3 Å, no methanol methyl groups are within 3.6 Å and no methanol hydroxyl groups are within 3.9 Å, of the center of the methane molecule. Methanol mole fractions are: 0.00 (solid circles); 0.05 (open circles); 0.10 (solid squares); 0.20 (open squares); 0.30 (solid triangles). Lines are gaussian fit to the data. The energy distributions are successively shifted upwards by 2 units for visual clarity.
APPENDIX C

Supplementary Data for Site Specific Co-solvent Preferential Interactions on Protein Surface through Quasi-Chemical Theory

Following the quasi-chemical generalization of potential distribution theorem [29, 213, 214], the excess chemical potential of a protein site in aqueous solution is defined as,

\[ \mu_{\text{ex}}^\alpha = kT \ln x_0 - kT \ln p_0 + kT \ln \langle e^{+\beta \epsilon} | n, m = 0 \rangle. \tag{C.1} \]

Here, \( x_0 \) is the probability of finding no water or co-solvents molecules within a defined proximal volume or inner shell around a protein site, \( \alpha \); \( p_0 \) is the probability of finding a cavity in bulk solution corresponding to the inner shell; and \( \ln \langle e^{+\beta \epsilon} | n, m = 0 \rangle \) is the outer shell contribution that accounts for the interaction of the protein site with the solvent in the outer shell under the condition that there is no solvent in the inner shell.

As the distribution of interaction energies of the investigated protein sites (Table 7.1) are nearly gaussian as shown in Fig. C.1, the rightmost term in Eq. C.1 simplifies to the sum of average interaction energy, \( \langle \epsilon \rangle \), and variance, \( \sigma^2 \), of the distribution of interaction energies,

\[ \mu_{\text{ex}}^\alpha = kT \ln x_0 - kT \ln p_0 + \langle \epsilon \rangle + \beta \sigma^2 / 2. \tag{C.2} \]
The individual QC contributions to excess chemical potential of specific proteins sites on lysozyme surface in pure water, 2 mol% urea solution and 10 mol% methanol solution are shown in Figs. C.2, C.3 and C.4, respectively. The co-solvent preferential interaction parameters for the protein sites calculated as derivative of $\mu^e_x$ with respect to bulk co-solvent concentration,

$$
\Gamma_{ac} = -\rho_c \left( \frac{\partial \ln x_0}{\partial \rho_c} \right)_{T,P,\rho_p} + \rho_c \left( \frac{\partial \ln \rho_0}{\partial \rho_c} \right)_{T,P,\rho_p} - \rho_c \left( \frac{\partial \ln \langle e^{+\beta \epsilon} | n, m = 0 \rangle}{\partial \rho_c} \right)_{T,P,\rho_p},
$$

and their corresponding QC contributions are given in Figs. C.5 and C.6 for aqueous solution of urea and methanol, respectively. The mean and standard deviation of the site-specific preferential interaction parameters calculated over different inner shell volumes, i.e., for $\lambda_w$ ranging from 2.6 to 3.0 Å, are collected in Table. C.1.
Figure C.1: Logarithm of conditional probability distribution of interaction energies of 7 protein sites (Table. 7.1) in neat water (top panel) and in aqueous solution of 2 mol% urea (bottom left panel) and 10 mol% methanol (bottom right panel) such that no water or co-solvent molecule is within inner shell volume defined by water conditioning radius, $\lambda_w$, of 2.8 Å. The corresponding conditioning radii for heavy atoms of urea and methanol are given in Table 7.3. Symbols: protein site 1 = black circles; protein site 2 = red circles; protein site 3 = green circles; protein site 4 = blue circles; protein site 5 = violet circles; protein site 6 = dark green circles; protein site 7 = orange circles. Lines are the gaussian fit to the simulation data. Successive energy distributions are shifted upwards by 2 units for visual clarity. The average interaction energy of the protein site 7 (carbonyl group of Leu129) is shifted towards right by 22 kcal/mol to fit all distribution within the same graph.
Figure C.2: Quasi chemical contributions to excess chemical potential of 7 protein sites in neat water as a function of water conditioning radius, $\lambda_w$. Symbols: $kT \ln \alpha_0 = $ solid circles; $-kT \ln p_0 = $ open circles; $\langle \epsilon \rangle = $ solid squares; $\beta \sigma^2/2 = $ open squares; $\mu^\alpha_{ex} = $ open triangles. Lines are drawn connecting the data points to guide the eye.
Figure C.3: Quasi chemical contributions to excess chemical potential of 7 protein sites in 2 mol% aqueous urea solution as a function of water conditioning radius, $\lambda_w$. The corresponding conditioning radii for the heavy atoms of urea are calculated as $\lambda_i = \lambda_w - 1.3 + R_i$. Symbols: $kT \ln x_0 = $ solid circles; $-kT \ln p_0 = $ open circles; $\langle \epsilon \rangle = $ solid squares; $\beta \sigma^2/2 = $ open squares; $\mu^{ex}_\alpha = $ open triangles. Lines are drawn connecting the data points to guide the eye.
Figure C.4: Quasi chemical contributions to excess chemical potential of 7 protein sites in 10 mol% aqueous methanol solution as a function of water conditioning radius, $\lambda_w$. The corresponding conditioning radii for the heavy atoms of methanol are calculated as $\lambda_i = \lambda_w - 1.3 + R_i$. Symbols: $kT \ln x_0 =$ solid circles; $-kT \ln p_0 =$ open circles; $\langle \epsilon \rangle =$ solid squares; $\beta \sigma^2/2 =$ open squares; $\mu_{\alpha}^{ex} =$ open triangles. Lines are drawn connecting the data points to guide the eye.
Figure C.5: Quasi chemical contributions to preferential interaction parameters of 7 protein sites in 2 mol% aqueous urea solution as a function of water conditioning radius, $\lambda_w$. The corresponding conditioning radii for the heavy atoms of urea are calculated as $\lambda_i = \lambda_w - 1.3 + R_i$. Symbols: $kT\ln x_0 = $ solid circles; $-kT\ln p_0 = $ open circles ; $\langle \epsilon \rangle = $ solid squares; $\beta \sigma^2/2 = $ open squares; $\Gamma_{ac} = $ open triangles. Lines are drawn connecting the data points to guide the eye.
Figure C.6: Quasi chemical contributions to preferential interaction parameters of 7 protein sites in 10 mol% aqueous methanol solution as a function of water conditioning radius, $\lambda_w$. The corresponding conditioning radii for the heavy atoms of methanol are calculated as $\lambda_i = \lambda_w - 1.3 + R_i$. Symbols: $kT \ln x_0 =$ solid circles; $-kT \ln p_0 =$ open circles; $\langle \epsilon \rangle = $ solid squares; $\beta_0^2/2 =$ open squares; $\Gamma_{\alpha c} =$ open triangles. Lines are drawn connecting the data points to guide the eye.
Table C.1: Inner shell and total preferential interaction parameters of 7 protein sites obtained through QC analysis in 2 mol% urea solution and 10 mol% methanol solution, averaged over five inner shell volumes i.e., for $\lambda_w$ ranging from 2.6 to 3.0 Å.

<table>
<thead>
<tr>
<th>Protein site index</th>
<th>$\Gamma_{ac}$ (Inner Shell)</th>
<th>$\Gamma_{ac}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2 mol% urea solution</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.061±0.018</td>
<td>0.267±0.063</td>
</tr>
<tr>
<td>2</td>
<td>0.060±0.007</td>
<td>0.206±0.065</td>
</tr>
<tr>
<td>3</td>
<td>0.064±0.011</td>
<td>0.101±0.026</td>
</tr>
<tr>
<td>4</td>
<td>0.055±0.020</td>
<td>0.595±0.040</td>
</tr>
<tr>
<td>5</td>
<td>0.050±0.022</td>
<td>0.639±0.028</td>
</tr>
<tr>
<td>6</td>
<td>-0.003±0.009</td>
<td>0.009±0.063</td>
</tr>
<tr>
<td>7</td>
<td>-0.085±0.028</td>
<td>0.083±0.102</td>
</tr>
<tr>
<td><strong>10 mol% methanol solution</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.153±0.063</td>
<td>0.708±0.083</td>
</tr>
<tr>
<td>2</td>
<td>0.149±0.083</td>
<td>0.670±0.039</td>
</tr>
<tr>
<td>3</td>
<td>0.104±0.017</td>
<td>-0.606±0.046</td>
</tr>
<tr>
<td>4</td>
<td>0.092±0.031</td>
<td>-0.003±0.130</td>
</tr>
<tr>
<td>5</td>
<td>0.131±0.074</td>
<td>1.080±0.064</td>
</tr>
<tr>
<td>6</td>
<td>0.135±0.075</td>
<td>0.873±0.078</td>
</tr>
<tr>
<td>7</td>
<td>0.206±0.054</td>
<td>-1.271±0.132</td>
</tr>
</tbody>
</table>
BIBLIOGRAPHY


