THE UNIVERSITY OF CHICAGO

MODELING THE HYDRATION LAYER AROUND PROTEINS AND NUCLEIC ACIDS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY

JOUKO J. VIRTANEN

CHICAGO, ILLINOIS
MARCH 2012
TO MY FAMILY
# Table of Contents

List of Figures................................................................................................................................................ix

List of Tables..................................................................................................................................................xiii

Acknowledgments..........................................................................................................................................xv

Abstract......................................................................................................................................................xvi

1. Hydration................................................................................................................................................1

1.1. Why study hydration?.........................................................................................................................1

1.2. Reconstructing hydration shells......................................................................................................1

1.3. The HyPred model...............................................................................................................................2

1.4. Applications........................................................................................................................................3

1.5. Outline.................................................................................................................................................4

2. Modeling the hydration layer around proteins: HyPred.................................................................8

2.1. Abstract..............................................................................................................................................8

2.2. Introduction.......................................................................................................................................9

2.3. Methods...........................................................................................................................................12

2.3.1. Molecular dynamics simulations..................................................................................................12

2.3.2. Modeling the density from MD simulations...............................................................................13

2.3.3. Reconstruction the density directly.............................................................................................16

2.4. Results and discussion....................................................................................................................18

2.4.1. Simulations and radial distribution functions.............................................................................18
4.2. Introduction........................................................................................................................................78

4.3. Results..................................................................................................................................................81

4.3.1. Solvent density maps.....................................................................................................................81

4.3.2. Comparison between A and B-forms............................................................................................83

4.3.3. Comparison between A-forms of RNA and DNA........................................................................86

4.3.4. High density regions.....................................................................................................................90

4.3.5. Proximal radial distribution functions........................................................................................91

4.3.6. Comparison to protein pRDFs......................................................................................................96

4.3.7. Reconstructions............................................................................................................................97

4.4. Discussion..........................................................................................................................................102

4.5. Methods............................................................................................................................................103

4.5.1. MD simulations............................................................................................................................103

4.5.2. HyPred..........................................................................................................................................105

4.5.3. Quantifying the quality of HyPred predictions............................................................................108

5. HyPred based electrostatic calculations.............................................................................................110

5.1. Abstract............................................................................................................................................110

5.2. Introduction....................................................................................................................................110

5.3. Methods..........................................................................................................................................112

5.3.1. MD simulations............................................................................................................................112

5.3.2. HyPred........................................................................................................................................114

5.3.3. Electrostatic energy calculations.................................................................................................116

5.3.4. Nonlinear Poisson-Boltzmann calculations...............................................................................117
5.4. Results

5.4.1. Solute/solvent electrostatic interaction energies from MD simulations

5.4.2. Charge density maps

5.4.3. Dependence of charge density maps on ionic conditions

5.4.4. Charge pREDFs

5.4.5. Dependence of pREDFs on ionic conditions

5.4.6. Comparison between protein and RNA ion pRDFs

5.4.7. Solute/solvent electrostatic interaction energies from charge density maps

5.4.8. Charge density maps predicted by HyPred

5.4.9. Solute/solvent electrostatic interaction energies calculated by HyPred

5.4.10. Comparison to DelPhi

5.5. Conclusion

6. Conclusion

6.1. Summary

6.2. Future directions

6.2.1. Binding

6.2.2. Orientations of water molecules

6.2.3. Free energy of hydration

6.2.4. Residency times

6.2.5. Hydrogen bonding

6.2.6. Crystallographic refinement

6.2.7. Improving the accuracy of HyPred
6.2.8. Effect of temperature and pressure.................................................................156

6.2.9. Different water models..................................................................................156

6.2.10. Development of an implicit solvent model.................................................157

References................................................................................................................159
List of Figures

Figure 2-1. Calculating solvent density around Ub using pRDFs......................................................15

Figure 2-2. Protein motions affect calculated solvent density..........................................................19

Figure 2-3. Specificity and transferability of pRDFs........................................................................21

Figure 2-4. HyPred reconstructions..................................................................................................23

Figure 2-5. Hydration density at 2 Å resolution.............................................................................31

Figure 2-6. Predicted water molecules around proteins....................................................................32

Figure 2-7. Water molecules around proteins..................................................................................34

Figure 3-1. Evidence of sufficient sampling.....................................................................................46

Figure 3-2. Equilibration....................................................................................................................47

Figure 3-3. Comparison between calculated and experimental SWAXS patterns.......................51

Figure 3-4. Cross section of the hydration shell electron densities of Ub, HEWL, and Mb as
determined by MD simulations and predicted by HyPred.............................................................52

Figure 3-5. The angle used in the enhanced version of HyPred.......................................................54

Figure 3-6. Comparison between calculated and experimental SWAXS patterns.......................56
Figure 3-7. The SWAXS patterns for concentrations of Mb ranging from 200 mg/mL to 4.52 mg/mL and the SWAXS pattern calculated from the MD simulation.................................57

Figure 3-8. Comparison between calculated and experimental SWAXS patterns for cytochrome c........................................................................................................58

Figure 3-9. Influence of dynamics on scattering profiles for Ub..................................................60

Figure 3-10. SWAXS patterns calculated from a crystal structure of Mb as well as from 4 different NMR structures of Mb..................................................................................62

Figure 3-11. Effect of discretizing the solvent..................................................................................64

Figure 3-12. The influence of discretizing and averaging the electron density into (0.5 Å)³ cubes..................................................................................................................65

Figure 3-13. Comparison between various hydration models and CRYSOL...............................68

Figure 3-14. Comparison between various hydration models..........................................................70

Figure 3-15. Agreement between the MD simulation with either the experimental data or the various hydration models.................................................................74

Figure 3-16. P(r) curves calculated using various data ranges..........................................................75

Figure 3-17. Comparison between a direct Fourier transformation of the scattering calculated from the MD simulation of HEWL out to q=30 Å⁻¹ and the P(r) curve calculated from the HEWL crystal structure hydrated by HyPred..............................76
Figure 4-1. Cross sections showing the hydration shell density of 1A1T at 20 mM NaCl, 145 mM NaCl, 145 mM NaCl and 10 mM MgCl$_2$, and 1 M NaCl ................................................................. 82

Figure 4-2. pRDFs calculated from MD simulations performed at 20 mM and 1 M NaCl........83

Figure 4-3. Comparison of the hydration of CG pairs between A-form and B-form nucleic acids and between DNA and RNA........................................................................................................................................84

Figure 4-4. Cross section of the same CG base pair as displayed in Figure 4-3, with the full RNA structure........................................................................................................................................85

Figure 4-5. Comparison of the hydration of UA pairs between A-form and B-form RNA and between TA pairs of A-form DNA and B-form DNA........................................................................................................88

Figure 4-6. More examples of cross sections of hydration shells........................................89

Figure 4-7. Comparison between the hydration shells obtained from MD simulations and predicted by HyPred.......................................................................................................................................91

Figure 4-8. pRDFs of RNA and protein atom types................................................................93

Figure 4-9. pRDFs categorized according to the concavity of the region in which the cube is located, the second nearest neighbor, and the angular dependence.........................................................95

Figure 4-10. A comparison of the hydration shells of an RNA molecule and protein as computed from MD simulations and by HyPred.................................................................................................101

Figure 5-1. Cross sections of the solvent charge density from MD simulations and HyPred predictions of the separate RNA and protein portions of 2A9X........................................................................125

Figure 5-2. Influence of ionic condition on charge density distribution around RNA.........127

Figure 5-3. Influence of ionic condition on charge density distribution around proteins ......128

Figure 5-4. The separate contributions of water and ions to the total charge density distribution at 20 mM and 1 M NaCl.........................................................................................................................129

Figure 5-5. Three dimensional distribution of high positive and negative charge densities...130
Figure 5-6. Charge density pREDFs for the CH and O atom types at 20 mM and 1 M NaCl.....132

Figure 5-7. The difference in the pREDFs for the O atom type at 1 M and 20 mM NaCl........133

Figure 5-8. The sodium and chloride ion pRDFs for the CH and O atom types from the 145 mM and 1 M NaCl MD simulations.................................................................135

Figure 5-9. Comparison of protein and RNA sodium and chloride ion pRDFs at 145 mM and 1 M NaCl........................................................................................................136

Figure 6-1. General pRDFs of the hydration shells of the separate protein and RNA portions of 1A1T and 1G70, along with general pRDFs of the regions of the hydration shells that are buried upon binding........................................................................................................149

Figure 6-2. The number of RNA water molecules with a protein water molecule within a distance of r................................................................................................................150

Figure 6-3. Orientational pRDFs.............................................................................................151

Figure 6-4. Residency times compared with density................................................................153

Figure 6-5: A comparison of the TIP3P and TIP4P water models........................................157
List of Tables

Table 2-1. The list of atom types used in the set of fine atom types
Table 2-2. Accuracy of the density reconstruction at 0.5 Å grid spacing
Table 3-1. The difference between the specified scattering intensity and that calculated from
the MD simulation, quantified as ζ
Table 3-2. Rg values
Table 4-1. The R factors for the HyPred predictions with respect to the MD simulations for the
RNA molecules
Table 5-1. The number of water molecules and ions in each simulation as well as the volume
in Å³
Table 5-2. Solute/solvent electrostatic interaction energies of RNA molecules calculated from
MD simulations
Table 5-3. Solute/solvent electrostatic interaction energies of proteins calculated from MD
simulations
Table 5-4. The charges on the protein and RNA molecules used in the simulations
Table 5-5: Comparison of the solute/solvent electrostatic interaction energies for RNA
molecules calculated from MD simulations, charge density maps obtained from MD
simulations, HyPred, and nonlinear Poisson-Boltzmann equation
Table 5-5: Comparison of the solute/solvent electrostatic interaction energies for proteins calculated from MD simulations, charge density maps obtained from MD simulations, HyPred, and nonlinear Poisson-Boltzmann equation.

Table 5-6: Comparison of the solute/solvent electrostatic interaction energies for proteins calculated from MD simulations, charge density maps obtained from MD simulations, HyPred, and nonlinear Poisson-Boltzmann equation.
Acknowledgments

This is of course the most emotional part of my thesis. There are so many people who have helped me get where I am today. First I would thank my parents, Jorma and Sinnika, my brother Jukka, and my sister Seija. I am lucky to have been born to such a loving and supportive family, which values educations as much as my family does. I am lucky to have attended schools with many great teachers. I would like to thank my undergraduate advisors Ara Apkarian and Richard Saykally. It has been an honor to work with my advisor, Karl Freed, who has given me a good mix of freedom and advice. Tobin Sosnick has an unlimited enthusiasm and energy for science and everything else. Lee Makowski has been a great help with SWAXS. Susan Karani supported me emotionally for the four years that we were together. Her family took me in and treated me as one of their own. I would like to thank fellow graduate students in the Freed and Sosnick labs. My office mate Aashish Adhikari has helped me on many occasions and I have enjoyed sharing an office with him. Marc Parisien, the RNA guru, helped a great deal while writing my paper on the hydration of nucleic acids. Esmael Haddadian has advised me on how to run MD simulations and how to use VMD. I would like to thank James Hinshaw, Joe DeBartolo, Glen Hockey, Heipeng Gong, and Abhishek Jha for numerous reasons. I would like to thank Laura Hawk, Christine Schwerdtfeger, and Britni Ratliff for many weekends spent working together on problem sets and listening to my candidacy exam presentation umpteen times.
Abstract

We develop a theory, termed HyPred, for describing the hydration shells of biological macromolecules and use the model to calculate accurate small and wide angle X-ray scattering (SWAXS), study protein-RNA binding, and calculate electrostatic energies of hydration. HyPred is developed using the approximation that a solute molecule’s hydration is determined by local properties and that long range interactions may be neglected. Thus, by examining the hydration of several different biological macromolecules, the hydration of other similar molecules (e.g., proteins, RNA, DNA) can be predicted because each class has similar surface chemical properties. More specifically, all-atom explicit solvent molecular dynamics simulations are performed for several proteins and nucleic acids and I calculate the average solvent density in each cube of a discretized representation of the simulation boxes. Proximal radial distribution functions (pRDFs) are calculated evaluating the average density of cubes as a function of distance from the nearest solute atom and a function of the atom type of the nearest solute atom. This model is validated by reproducing the solvent densities found in the MD simulations and comparing the locations of predicted water molecules to those in the crystal structures. The HyPred model is further validated by comparing calculated SWAXS patterns of proteins, hydrated by HyPred, to experimental data. The HyPred model is improved by the addition of the dependence of the pRDFs on the second nearest neighbor, orientation, and the local geometry of the solute. HyPred is then extended to model the hydration layer of
nucleic acids. Finally, HyPred is extended to predict the charge density and the electrostatic energies of hydration and electrostatic potential energy maps.
1. Hydration

1.1 Why study hydration?

The human body contains approximately 57% water, with variations from person to person and stage of life(1). Water is necessary for all life on Earth. The hydrophobic effect is the primary cause of protein folding(2, 3), and water has been implicated in smoothing the protein folding landscape(4). Proteins fold in order to hide their hydrophobic portions from water. Frauenfelder et al. argue that protein dynamics are slaved to solvent dynamics(5). This thesis deals with the hydration of protein and nucleic acid molecules. Hydration is essential to the functions of proteins and RNA molecules. Proteins are involved in signaling, act as catalysts, and carry out molecular transport, while RNA molecules are involved with the translation of DNA into proteins. mRNAs temporarily store genetic information before being translated into proteins, spliceosomes splice mRNAs, such that many different proteins can be formed from bits and pieces of the same DNA sequence, vastly increasing the complexity of living organisms, and ribosomes translate mRNA into proteins.

1.2 Reconstructing Hydration Shells

There are many applications for which the computation of a biological macromolecule’s hydration properties are desired but for which performing an all atom MD simulations is impractical, for example, if the hydration must be studied for a large number of protein ligand binding partners, so that long all-atom explicit solvent MD simulations are not feasible.
Therefore, researchers have worked on fast methods for predicting a molecule’s hydration. These methods generally fall into one of three categories. The first involves the use of modified radial distribution functions, such as pRDFs (6-9). The second uses statistical mechanical methods such as 3D-RISM and integral equations(10). The statistical mechanical methods are slower(11), then those based on pRDFs. The third approach uses the locations of waters in crystallographic structures(12-15). However, even high resolution crystal structures do not present a completely accurate or complete picture of hydration. HyPred, the model for predicting the hydration around proteins and nucleic acids introduced here, falls into the first category. HyPred introduces several improvements over previous methods(6, 16). These improvements involve including more detailed specification of atom types, using the van der Waals radius and partial charge of the atoms in determining the cube assignments, and the dependence on angular orientations of bonding, second nearest neighbor, and the local geometry of the surface of the macromolecule(17).

1.3 The HyPred model

HyPred uses information from MD simulations to predict the hydration shell densities of macromolecules for which MD simulations have not been performed. MD simulations have been carried out for several proteins, RNA molecules, and 2 DNA molecules. Solvent density maps are calculated from these MD simulations and proximal radial distribution functions (pRDFs) are calculated from these density maps. For each cube, the distance to the nearest atom surface is computed, where the atom surface is formed by the van der Waals radius
multiplied by a scale factor plus the absolute value of the charge multiplied by a scale factor (i.e. \( r = v_{\text{vdw}}a + |q|b \)). Cubes which belong to the same atom type and lie within the same distance range from the solute surface are grouped together and their densities are averaged to obtain the value of the pRDF at that distance for that atom type. This process can be reversed to reconstruct the hydration shells of the simulated macromolecules and to predict the hydration layers of solute molecules that have not been simulated. In addition to the atom type of the nearest solute atom and distance to that atom, HyPred takes into account the dependence upon the second nearest neighbor, the angle formed by the vector connecting the cube and the nearest solute atom and the vector connecting the nearest solute atom and the atom which it is bonded, and the local geometry of the surface of the macromolecule.

1.4 Applications

SWAXS is an important experimental technique which is often used to validate or invalidate proposed models of macromolecule structures, to obtain envelopes of macromolecules (18), or interparticle distance distributions (19). The calculation of SWAXS patterns is complicated by the need to include solvent effects. The first solvent effect is contrast. That is, the SWAXS pattern is determined by all electron to electron distances within the system, including solute-solvent and solvent-solvent distances. The second solvation effect is the presence of a variable density throughout the hydration shell. The standard program for calculating SWAXS patterns is CRY SOL. CRY SOL assumes the presence of a uniform hydration layer which is 10% greater in density than bulk solvent. Often times the density of the
hydration shell is adjusted so that the calculated SWAXS pattern matches experimental data. HyPred eliminates the need to optimize phenomenological parameters and greatly improves the fit to experimental data.

HyPred has been adopted to predict solvent charge densities as well as solvent electron densities. The charge density around a macromolecule can be used to calculate the electrostatic energy of solvation. Pettitt has also calculated electrostatic energies of solvation, but not with a water model as detailed as HyPred. The Poisson-Boltzmann equation can be used to calculate electrostatic free energies of solvation. If the electrostatic energies are calculated for a protein, an RNA molecule, and the protein and RNA bound together, then the electrostatic energy of binding can be calculated. pRDFs have been developed for different ionic conditions, so that HyPred is able to predict charge and electron densities at various ionic concentrations.

1.5 Outline

In Chapter 2, the HyPred theory is developed to describe the hydration of proteins. The HyPred model is based on the approximation that the density of water at a particular location in the hydration shell is only determined by the local environment and not long range interactions. Thus, by examining the hydration shells of proteins and grouping regions with similar characteristics in those hydration shells, the hydration shells of other molecules with similar characteristics can be predicted accurately. As a feasibility test, HyPred is used to reconstruct the hydration shells of the same proteins whose MD simulations are used to
develop HyPred, and three metrics are used to quantify the ability of HyPred to reconstruct the hydration shells from the MD simulations. Next, HyPred is used to predict the hydration shell densities of proteins other than the proteins that were used to develop HyPred, and the qualities of these predictions are quantified as before. The HyPred model is used to predict the locations of crystallographic water molecules, and the accuracy of the prediction is quantified by calculating the distances between the crystallographic water molecules and the predicted water molecules. Two sets of pRDFs are tested. In the first coarse grained set, heavy atoms are grouped by element, whereas hydrogen atoms are further subdivided according to the element to which they are bonded. In the second fine grained set, every unique atom is its own type.

In chapter 3, the HyPred model is applied to the calculation of Small/Wide angle X-ray scattering (SWAXS). I develop and test various methods of calculating SWAXS patterns. In the first method, SWAXS patterns are computed directly from MD simulations. In the second method, the electrons densities in the hydration shells of the MD simulations are averaged and discretized into cubes. The SWAXS intensities are calculated from the protein and averaged hydration shell densities. Finally, HyPred is used to predict the electron densities in the cubes, and the SWAXS pattern is calculated from the HyPred hydrated model. The different calculations are compared with experimental data, to each other, and to CRYSOL, the current standard for SWAXS calculations. I investigate deviations between the various computational methods and experiment. I explore the effects of protein dynamics on the SWAXS pattern of ubiquitin.
In Chapter 4, the HyPred model is extended to RNA molecules. I examine the hydration patterns around several different RNA molecules and analyze the hydration patterns around different base pairs. pRDFs are calculated for the different RNA atom types. pRDFs from simulations of a given RNA are used first to predict its hydration layer, and then the pRDFs of other RNA molecules are used to predict its hydration. The qualities of these reconstructions are quantified as before.

In Chapter 5, HyPred is extended to predict charge densities in addition to electron densities. This extension enables the calculation of solute/solvent electrostatic interaction energies \( E_{\text{solute/solvent}} \) and electrostatic potential energy maps. Solute/solvent electrostatic interaction energies are calculated from MD simulations of proteins and RNA molecules at four different ionic conditions. Charge density maps are generated from the MD simulations and analyzed. pREDFs, the charge density analogs of pRDFs, are evaluated from the charge density maps and examined in detail. The pREDFs are used to predict the charge density around the proteins and nucleic acids in the same manner as pRDFs are used to calculate the electron density around solute molecules. The predicted charge density maps are compared to the charge density maps obtained from the MD simulations. Electrostatic potential energy maps are calculated from the HyPred predicted charge density maps. Solute/solvent electrostatic interaction energies are calculated from the charge density maps predicted by HyPred and compared to the values provided by the MD simulations. The nonlinear Poisson-Boltzmann equation is solved using DelPhi for the proteins and RNA molecules under study, \( E_{\text{solute/solvent}} \) is evaluated, and compared to the results from the MD simulations and HyPred predictions. For
each type of calculation the influence of the ionic conditions is assess and the contributions of water and ions are examined separately.

Chapter 6 discusses future work, including the introduction of even more detailed pRDFs that contain information about the distribution of water orientations, residency times, water-water hydrogen bonding. New applications could include binding, calculating free energies of hydration, crystallographic refinement, and the development of implicit solvent models. HyPred could be improved by studying the effect of temperature on the hydration. This would also provide insight into the temperature dependence of the hydrophobic effect. A better understanding of the role of surface geometry on hydration could be used to improve HyPred.
2. Modeling the hydration layer around proteins: HyPred

2.1 Abstract

The contents of this chapter have been published(17). Protein hydration plays an integral role in determining protein function and stability. We develop a simple method with atomic level precision for predicting the solvent density near the surface of a protein. A set of proximal radial distribution functions are defined and calculated for a series of different atom types in proteins using all-atom, explicit solvent molecular dynamic simulations for three globular proteins. A major improvement in predicting the hydration layer is found when the protein is held immobile during the simulations. The distribution functions are used to develop a model for predicting the hydration layer with sub 1 Å resolution without the need for additional simulations. The model and the distribution functions for a given protein are tested in their ability to reproduce the hydration layer from the simulations for that protein as well as those for other proteins. Predictions for the density of water in the hydration shells are then compared with high occupancy sites observed in crystal structures. The accuracy of both tests demonstrates that the solvation model provides a basis for quantitatively understanding protein solvation and thereby predicting the hydration layer without additional simulations.
2.2 Introduction

The solvent affects the thermodynamics and kinetics of numerous biological processes, including protein and nucleic acid folding, stability (20, 21) and dynamics (22), enzymology, including transition state stabilization (23), binding (24, 25), diffusion, electrostatic interactions (26), charge transfer reactions, ion channel and membrane transporter conductance (27), etc. It is difficult to think of a biological process that is independent of solvation. In addition, the presence of a hydration layer surrounding proteins influences many biophysical measurements, including NMR spectroscopy (28-30), X-ray crystallography (31, 32), small and wide angle X-ray scattering (SWAXS) (33), and neutron diffraction (34, 35). The interpretation of data from all these applications would benefit by having a rapid and accurate model to predict the solvent density around biomolecules. Moreover, the model would provide a fundamental physical basis for describing solvent-biomolecule interactions.

The hydration model we advance here extends and refines the strategy of Pettitt and coworkers of using MD simulation data to predict the hydration shell densities surrounding proteins and DNA (6, 7, 36, 37). The hydration layers deduced from the simulations are converted to a set of proton proximal radial distribution functions (pRDFs) for different atoms types (e.g., N, C and O). By proton we mean the eight protons of the oxygen atoms and the protons of the hydrogen atoms. These distribution functions describe the solvent density at a position located at a distance $r$ from the closest solute atom which has the designation of atom type $a$. Subsequent inversion of this mapping process generates a predicted hydration shell
density around a protein. This methodology provides the important possibility of predicting the hydration layer of any soluble protein without additional MD simulations.

We address an apparently minor deficiency of Pettitt and coworkers’ approach and find a substantial improvement in the prediction of the hydration layer. Our MD simulations maintain the protein atoms immobile, and only the water molecules are allowed to move, whereas their protein is mobile. Additionally, their procedure only specifies different pRDFs for a few distinct atom types, specifically, one each for O, N and C, while ignoring all solute hydrogen atoms (36). Nevertheless, they recognized the benefit of calculating the pRDFs for additional atom types (37), and did use many atom types when examining DNA(16). We include solute hydrogen atoms and further categorize the atoms into subclasses (38) depending on the individual position within each amino acid to generate a total of ~300 distinct pRDFs (Table 2-1). Other improvements in our study include the use of longer simulations and finer grid spacing to achieve higher resolution (0.5 versus 2 Å). Finally, we demonstrate the transferability of the pRDFs by predicting the hydration shell density of a protein using pRDFs from simulations for other proteins. This result indicates the existence of a universal set of pRDFs for describing the hydration layer around globular proteins. Further tests involve the comparison of the predicted hydration layers with those observed in X-ray crystal structures.
Table 2-1: The list of atom types used in the set of fine atom types. Continued on next page.

<table>
<thead>
<tr>
<th>Atom Type 1</th>
<th>Atom Type 2</th>
<th>Atom Type 3</th>
<th>Atom Type 4</th>
<th>Atom Type 5</th>
<th>Atom Type 6</th>
<th>Atom Type 7</th>
<th>Atom Type 8</th>
<th>Atom Type 9</th>
<th>Atom Type 10</th>
<th>Atom Type 11</th>
<th>Atom Type 12</th>
<th>Atom Type 13</th>
<th>Atom Type 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALA N</td>
<td>CYS HBx</td>
<td>HEM CBx</td>
<td>HSP O</td>
<td>PHE N</td>
<td>TRP CG</td>
<td>GUA O4'</td>
<td>URA O3'</td>
<td>CYT H42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALA HN</td>
<td>CYS SG</td>
<td>HEM CGx</td>
<td>ILE N</td>
<td>PHE HN</td>
<td>TRP CD1</td>
<td>GUA C1'</td>
<td>ADE P</td>
<td>CYT N3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALA CA</td>
<td>CYS HG1</td>
<td>HEM O1x</td>
<td>ILE HN</td>
<td>PHE CA</td>
<td>TRP HD1</td>
<td>GUA H1'</td>
<td>ADE O1P</td>
<td>CYT C2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALA HA</td>
<td>CYS C</td>
<td>HEM O2x</td>
<td>ILE CA</td>
<td>PHE HA</td>
<td>TRP NE1</td>
<td>GUA N9</td>
<td>ADE O2P</td>
<td>CYT O2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALA CB</td>
<td>CYS O</td>
<td>HSD N</td>
<td>ILE HA</td>
<td>PHE CB</td>
<td>TRP HE1</td>
<td>GUA C8</td>
<td>ADE O5'</td>
<td>CYT C3'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALA HB</td>
<td>GLU N</td>
<td>HSD HN</td>
<td>ILE CB</td>
<td>PHE Hbx</td>
<td>TRP CE2</td>
<td>GUA H8</td>
<td>ADE C5'</td>
<td>CYT C3'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALA C</td>
<td>GLU HN</td>
<td>HSD CA</td>
<td>ILE HB</td>
<td>PHE CG</td>
<td>TRP CD2</td>
<td>GUA N7</td>
<td>ADE H5'</td>
<td>CYT H3'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALA O</td>
<td>GLU CA</td>
<td>HSD HA</td>
<td>ILE CGx</td>
<td>PHE CDx</td>
<td>TRP CE3</td>
<td>GUA C5</td>
<td>ADE H5''</td>
<td>CYT H2''</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG N</td>
<td>GLU HA</td>
<td>HSD CB</td>
<td>ILE HGx</td>
<td>PHE HDx</td>
<td>TRP HE3</td>
<td>GUA C6</td>
<td>ADE C4'</td>
<td>CYT O2'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG HN</td>
<td>GLU CB</td>
<td>HSD HBx</td>
<td>ILE CD</td>
<td>PHE CEx</td>
<td>TRP CZ3</td>
<td>GUA O6</td>
<td>ADE H4'</td>
<td>CYT H2'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG CA</td>
<td>GLU HB</td>
<td>HSD ND1</td>
<td>ILE Hdx</td>
<td>PHE HEx</td>
<td>TRP HZ3</td>
<td>GUA N1</td>
<td>ADE O4'</td>
<td>CYT O3'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG HA</td>
<td>GLU CG</td>
<td>HSD HD1</td>
<td>ILE C</td>
<td>PHE C2x</td>
<td>TRP CZ2</td>
<td>GUA H1</td>
<td>ADE C1'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG CB</td>
<td>GLU HG</td>
<td>HSD CG</td>
<td>ILE O</td>
<td>PHE H2x</td>
<td>TRP HZ2</td>
<td>GUA C2</td>
<td>ADE H1'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG HB</td>
<td>GLU CD</td>
<td>HSD CE1</td>
<td>LEU N</td>
<td>PHE C</td>
<td>TRP CH2</td>
<td>GUA N2</td>
<td>ADE N9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG CG</td>
<td>GLU OEx</td>
<td>HSD HE1</td>
<td>LEU HN</td>
<td>PHE O</td>
<td>TRP HH2</td>
<td>GUA H21</td>
<td>ADE C8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG HG</td>
<td>GLU H</td>
<td>HSD NE2</td>
<td>LEU CA</td>
<td>PRO N</td>
<td>TRP C</td>
<td>GUA H22</td>
<td>ADE H8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG CD</td>
<td>GLU O</td>
<td>HSD CD</td>
<td>LEU HA</td>
<td>PRO CD</td>
<td>TRP O</td>
<td>GUA N3</td>
<td>ADE N7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG HD</td>
<td>GLN N</td>
<td>HSD HD2</td>
<td>LEU CB</td>
<td>PRO Hdx</td>
<td>TYR N</td>
<td>GUA C4</td>
<td>ADE C5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG NE</td>
<td>GLN HN</td>
<td>HSD C</td>
<td>LEU HBx</td>
<td>PRO CA</td>
<td>TYR HN</td>
<td>GUA C3'</td>
<td>ADE C6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG HE</td>
<td>GLN CA</td>
<td>HSD O</td>
<td>LEU CG</td>
<td>PRO HA</td>
<td>TYR CA</td>
<td>GUA H3'</td>
<td>ADE N6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG C2</td>
<td>GLN HA</td>
<td>HSE N</td>
<td>LEU HG</td>
<td>PRO CB</td>
<td>TYR HA</td>
<td>GUA C2'</td>
<td>ADE H61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG NHx</td>
<td>GLN CB</td>
<td>HSE HN</td>
<td>LEU CDx</td>
<td>PRO HBx</td>
<td>TYR CB</td>
<td>GUA H2''</td>
<td>ADE H62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG HH</td>
<td>GLN HB</td>
<td>HSE CA</td>
<td>LEU Hdx</td>
<td>PRO CG</td>
<td>TYR Hbx</td>
<td>GUA O2'</td>
<td>ADE N1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG C</td>
<td>GLN CG</td>
<td>HSE HA</td>
<td>LEU C</td>
<td>PRO Hgx</td>
<td>TYR CG</td>
<td>GUA H2'</td>
<td>ADE C2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARG O</td>
<td>GLN HG</td>
<td>HSE CB</td>
<td>LEU O</td>
<td>PRO C</td>
<td>TYR CDx</td>
<td>GUA O3'</td>
<td>ADE H2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASN N</td>
<td>GLN CD</td>
<td>HSE HBx</td>
<td>LYS N</td>
<td>PRO O</td>
<td>TYR HDx</td>
<td>URA P</td>
<td>ADE N3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASN HN</td>
<td>GLN OEx</td>
<td>HSE ND</td>
<td>LYS HN</td>
<td>SER N</td>
<td>TYR CEx</td>
<td>URA O1P</td>
<td>ADE C4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASN CA</td>
<td>GLN NE2</td>
<td>HSE HD1</td>
<td>LYS CA</td>
<td>SER HN</td>
<td>TYR HEx</td>
<td>URA O2P</td>
<td>ADE C3'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASN HA</td>
<td>GLN HE</td>
<td>HSE CG</td>
<td>LYS HA</td>
<td>SER CA</td>
<td>TYR C2</td>
<td>URA O5'</td>
<td>ADE H3'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASN CB</td>
<td>GLN C</td>
<td>HSE CE1</td>
<td>LYS CB</td>
<td>SER HA</td>
<td>TYR OH</td>
<td>URA C5'</td>
<td>ADE C2'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASN HB</td>
<td>GLN O</td>
<td>HSE HE1</td>
<td>LYS HBx</td>
<td>SER CB</td>
<td>TYR HH</td>
<td>URA H5'</td>
<td>ADE H2''</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASN CG</td>
<td>GLY N</td>
<td>HSE NE</td>
<td>LYS CG</td>
<td>SER HBx</td>
<td>TYR C</td>
<td>URA H5''</td>
<td>ADE O2'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASN OD1</td>
<td>GLY HN</td>
<td>HSE HE2</td>
<td>LYS HGx</td>
<td>SER OG</td>
<td>TYR O</td>
<td>URA C4'</td>
<td>ADE H2'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASN ND2</td>
<td>GLY CA</td>
<td>HSE CD2</td>
<td>LYS CD</td>
<td>SER HG1</td>
<td>VAL N</td>
<td>URA H4'</td>
<td>ADE O3'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASN HD</td>
<td>GLY HA</td>
<td>HSE HD2</td>
<td>LYS HDx</td>
<td>SER C</td>
<td>VAL HN</td>
<td>URA O4'</td>
<td>CYT P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2-1 (continued)

<table>
<thead>
<tr>
<th>ASN C</th>
<th>GLY C</th>
<th>HSE C</th>
<th>LYS CE</th>
<th>SER O</th>
<th>VAL CA</th>
<th>URA C1'</th>
<th>CYT O1P</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASN O</td>
<td>GLY O</td>
<td>HSE O</td>
<td>LYS HEx</td>
<td>THR N</td>
<td>VAL HA</td>
<td>URA H1'</td>
<td>CYT O2P</td>
</tr>
<tr>
<td>ASP N</td>
<td>HEM FE</td>
<td>HSP C</td>
<td>LYS NZ</td>
<td>THR HN</td>
<td>VAL CB</td>
<td>URA N1</td>
<td>CYT O5'</td>
</tr>
<tr>
<td>ASP HN</td>
<td>HEM NA</td>
<td>HSP N</td>
<td>LYS HZx</td>
<td>THR CA</td>
<td>VAL HB</td>
<td>URA C6</td>
<td>CYT C5'</td>
</tr>
<tr>
<td>ASP CA</td>
<td>HEM NB</td>
<td>HSP HN</td>
<td>LYS C</td>
<td>THR HA</td>
<td>VAL CGx</td>
<td>URA H6</td>
<td>CYT H5'</td>
</tr>
<tr>
<td>ASP HA</td>
<td>HEM NC</td>
<td>HSP CA</td>
<td>LYS O</td>
<td>THR CB</td>
<td>VAL HGx</td>
<td>URA C5</td>
<td>CYT H5''</td>
</tr>
<tr>
<td>ASP CB</td>
<td>HEM ND</td>
<td>HSP HA</td>
<td>MET N</td>
<td>THR HB</td>
<td>VAL C</td>
<td>URA H5</td>
<td>CYT C4'</td>
</tr>
<tr>
<td>ASP HB</td>
<td>HEM C1x</td>
<td>HSP CB</td>
<td>MET HN</td>
<td>THR OGx</td>
<td>VAL O</td>
<td>URA C4</td>
<td>CYT H4'</td>
</tr>
<tr>
<td>ASP CG</td>
<td>HEM C2x</td>
<td>HSP HBx</td>
<td>MET CA</td>
<td>THR HG1x</td>
<td>X HTx</td>
<td>URA O4</td>
<td>CYT O4'</td>
</tr>
<tr>
<td>ASP ODx</td>
<td>HEM C3x</td>
<td>HSP CD</td>
<td>MET HA</td>
<td>THR CG2</td>
<td>X OTx</td>
<td>URA N3</td>
<td>CYT C1'</td>
</tr>
<tr>
<td>ASP ND</td>
<td>HEM C4x</td>
<td>HSP HD2</td>
<td>MET CB</td>
<td>THR HG2x</td>
<td>GUA P</td>
<td>URA H3</td>
<td>CYT H1'</td>
</tr>
<tr>
<td>ASP HD</td>
<td>HEM CHx</td>
<td>HSP CG</td>
<td>MET Hbx</td>
<td>THR C</td>
<td>GUA O1P</td>
<td>URA C2</td>
<td>CYT N1</td>
</tr>
<tr>
<td>ASP C</td>
<td>HEM HA</td>
<td>HSP NE</td>
<td>MET CG</td>
<td>THR O</td>
<td>GUA O2P</td>
<td>URA O2</td>
<td>CYT C6</td>
</tr>
<tr>
<td>ASP O</td>
<td>HEM HB</td>
<td>HSP HE2</td>
<td>MET HGx</td>
<td>TRP N</td>
<td>GUA O5'</td>
<td>URA C3'</td>
<td>CYT H6</td>
</tr>
<tr>
<td>CYS N</td>
<td>HEM HC</td>
<td>HSP ND</td>
<td>MET SD</td>
<td>TRP HN</td>
<td>GUA C5'</td>
<td>URA H3'</td>
<td>CYT C5</td>
</tr>
<tr>
<td>CYS HN</td>
<td>HEM HD</td>
<td>HSP HD1</td>
<td>MET CE</td>
<td>TRP CA</td>
<td>GUA H5'</td>
<td>URA C2'</td>
<td>CYT H5</td>
</tr>
<tr>
<td>CYS CA</td>
<td>HEM CMx</td>
<td>HSP CE</td>
<td>MET HEx</td>
<td>TRP HA</td>
<td>GUA H5''</td>
<td>URA H2''</td>
<td>CYT C4</td>
</tr>
<tr>
<td>CYS HA</td>
<td>HEM HMx</td>
<td>HSP HE1</td>
<td>MET C</td>
<td>TRP CB</td>
<td>GUA C4'</td>
<td>URA O2'</td>
<td>CYT N4</td>
</tr>
<tr>
<td>CYS CB</td>
<td>HEM CaX</td>
<td>HSP C</td>
<td>MET O</td>
<td>TRP HBx</td>
<td>GUA H4'</td>
<td>URA H2'</td>
<td>CYT H41</td>
</tr>
</tbody>
</table>

2.3 Methods

2.3.1 Molecular Dynamics Simulations

All-atom explicit solvent MD simulations have been performed at the temperature of 4°C for Ub (1UBQ (39); 6 ns), HEWL (6LYZ (40), 3 ns), Mb (1WLA(41), 6 ns), employing NAMD (42) and the CHARMM 27 all-atom force field (43). The proteins are solvated in a rectangular periodic box containing rigid TIP3P water molecules (44) and having dimensions 108 x 91 x 104 Å³ for Ub, 126 x 110 x 104 Å³ for lysozyme, and 112 x 120 x 99 Å³ for Mb. The ample simulation box dimensions are chosen for future applications and are truncated at 10 Å from the protein’s...
surface in order to speed up the calculations. Counterions are added to compensate for the charges on the proteins. The solvent for the Ub simulation consists of seven hydrogen phosphate ions, seven dihydrogen phosphate ions, 21 sodium ions, and 33976 TIP3P water molecules. HEWL is solvated by a solution with 16 acetate ions, eight sodium ions, and 47470 TIP3P water molecules. The Mb buffer contains five hydrogen phosphate ions, five dihydrogen phosphate ions, and 43828 TIP3P water molecules.

Energy minimization and equilibration proceeds in several stages. The solvent and protein hydrogen atoms are first energy minimized for 2000 steps, then, with the heavy atoms in the protein fixed in place, the temperature of each system begins at 1200 K, the systems are cooled to 4 °C over a period of 100 ps, and the systems are then equilibrated for 100 ps at 4 °C. The hydrogen atoms in the protein are fixed in place, and the systems are equilibrated for another 100 ps. All protein atoms remain immobile throughout the course of the subsequent simulation to render the comparison of the reconstructed density to the density of MD simulation more meaningful. Electrostatic interactions are computed with particle mesh Ewald summations. A 1 fs time step is used, and snapshots are saved every 1 ps. The simulation uses NVT conditions.

2.3.2 Modeling the density from MD simulations

We evaluate solvent densities in the first few hydration layers and use these densities to generate a model for hydrated proteins without the need for running additional computationally expensive MD simulations. The model is constructed from the average over
the MD simulation of the solvent’s proton density profile surrounding the protein. Figures 2-1a and b illustrate the method of calculating the pRDFs. The density is evaluated at 1 ps intervals for every (0.5 Å)³ cube situated outside the protein. Each cube exterior to the protein is assigned to an atom on the protein surface whose scaled van der Waals surface is closest to the center of the cube. An important difference between Pettitt’s work and the work presented here is that here the distance to the scaled van der Waals surface is used instead of the distance to the nuclei of the atom. The importance of this can be seen by considering a point that is within the van der Waals radius of a large atom but is closer to and lies outside of the van der Waals radius of a smaller atom. Pettitt’s method would predict the cube to contain some solvent density even though the space is insufficient for placing a solvent molecule. Thus, the new approach improves the density reconstruction exterior to the van der Waals radii. Since all protein atoms remain stationary during the simulations, the assignment of cubes to protein atoms only needs specification once at the beginning of the analysis.
**Figure 2-1: Calculating solvent density around Ub using pRDFs.** A, B) The average solvent density around Ub is obtained from MD simulations with the protein atoms fixed. An 8 Å grid spacing is shown in A) and C), but 0.5 Å is used in the calculations. C,D) Using these data, the pRDFs are calculated by identifying the nearest solute atom and distance to each grid element; for example, the oxygen atom (red, upper left) is the closest solute atom to 4 grid elements (denoted with lines). E,F) Solvent density calculated by reversing the mapping protocol using the pRDF calculated for the fine and coarse atom type definitions. R-values are listed. The color scale is asymmetric, and hence, noise tends to make the bulk solution appear blue. G) Solvent density calculated using the averaged pRDFs obtained from Mb and HEWL using the fine atom-type definition. Only one layer of cubes is shown but protein atoms within a 3 Å of the layer are shown. Thus some proteins atoms can be seen above the slab.
Protein atoms are grouped into classes using two different sets of atom types in order to test the degree of specification required for accurate reconstruction of the hydration layer. One specification collects heavy atoms into groups according to their element type (e.g., C, N, O, S...), while hydrogen atoms are grouped together depending on the atom to which they are bonded (e.g., CH, NH, OH...). The other more detailed specification defines the atom groups according to both their elemental character and amino acid type. The more detailed set assigns each unique atom in each amino acid type as an atom type (Table 2-1). The detailed set provides one source of improvement over Pettitt who classifies atoms only according to the element. When cubes are equidistant from atoms of the same type, the densities of the cubes are averaged according to eq 1. Illustrating this process for the atom type CH yields

\[ g_{CH}(r) = \frac{1}{N} \sum_{i=1}^{N} \rho_i \]

where \( \rho_i \) is the density of cube \( C_i \), the summation is performed over all cubes that are assigned to H atoms of type CH, the cubes lie at a distance between \( r - \delta r \) and \( r + \delta r \) from the protein’s H atom, and \( N \) is the number of cubes in the summation. This procedure provides \( g_{CH}(r) \), the proximal radial distribution function (pRDF), which can only be obtained by discretizing the simulation box into cubes.

### 2.3.3 Reconstructing the density directly

The reconstruction of the hydration shell density without additional MD simulations begins with the protein in the absence of water. The protein and surroundings are partitioned
into a grid of cubes as in the mean-field approach of the previous subsection. For each of the cubes the distance to the closest scaled van der Waals surface is calculated. \( R \) is the distance between the atom \( A \) and the center of cube \( i \). Each cube \( i \) outside of the protein is assigned the density \( g_A(r) \) from the pRDF for the given protein atom type, the separation \( r \), and the atom type closest to the cube’s center. The scale was optimized by minimizing the sum of the \( R \) factors of the three proteins. The optimum scale factor was found to be 0.53. Densities in cavities are set to zero. Assigning the densities for ubiquitin takes about 20-50 seconds. Thus, by determining the pRDFs for a single protein or an average for several proteins, the pRDFs can be used to evaluate the solvent electron density for other proteins without the need for additional MD simulations. We call this process of hydrating the protein “HyPred”.

The first test of the methods involves first reconstructing the hydration shell density of each protein using the pRDFs determined for that protein. Then, the hydration shell of each protein is evaluated with the average of the pRDFs of the other three proteins. When insufficient data are available for constructing the atom type pRDFs, the missing data are replaced by portions of the pRDFs of the elements.

Crystallographic water molecules are predicted from both the MD simulations and HyPred reconstructions. Water molecules are predicted at positions where the solvent density in a \((0.5 \, \text{Å})^3\) cube is above a threshold level except when another cube with a higher density lies within 2.8 Å (the diameter of a water molecule).
2.4 Results and Discussion

2.4.1 Simulations and radial distribution functions

We begin by performing all-atom MD simulations using explicit (TIP3P) solvent (45) for ubiquitin (Ub, 1UBQ (39)), hen egg white lysozyme (HEWL, 6LYZ (40)), and myoglobin (Mb, 1WLA (41)) and allowing only the solvent molecules to move (Fig. 2-1A). The solvent proton density is calculated for individual frames taken every 1 ps and is averaged over 3000-6000 frames. The local solvent density varies, but generally the simulations display a thin depletion layer just outside the protein (blue in Fig. 2-2B). The first hydration layer (red) is found ~ 1-2 Å from the protein’s surface and is followed by a region of reduced solvent density.

For comparison, we also calculate the hydration layer when the protein atoms are allowed to move during the simulations (Fig. 2-2A). Protein motions, particularly those of surface side chains, can be substantial (e.g., rms displacements of 3 Å). Rather than increasing the accuracy of the reconstruction of the hydration layer from a theoretical model, such movement, in fact, artifactually reduces the time-averaged solvent density near the protein as compared to the stationary protein case. For example, a cube, might be accessible to water molecules in one snapshot, but could become blocked to solvent in another snapshot due to a side chain motion. This inaccessibility reduces the average solvent density assigned to the cube, but the decrease is due solely to the physical presence of the side chain and not to an actual repulsion arising from proximity to nearby hydrophobic groups. Hence, permitting protein motions in the MD simulations leads to a gross overestimate of the size and magnitude of the
depletion layer at the surface of the protein and thereby impedes the accurate calculation of the reconstructions (Fig. 2-2). When the protein atoms are immobile, on the other hand, a significant but thinner solvent depletion layer remains around the entire protein, while a greater number of regions of high solvent density are observed than when the protein is allowed to move.

![Figure 2-2](image)

**Figure 2-2. Protein motions affect calculated solvent density.** A) Average solvent density obtained from MD simulations where the protein atoms are allowed to move and B) the resulting reconstruction. C,D) Reconstructions for two different protein conformations calculated using pRDFs obtained from the dynamic protein.
The peaks of maximum density in the pRDFs for hydrogens attached to the polar atoms, oxygen and nitrogen exceed those for hydrogen atoms attached to the more hydrophobic carbon atoms (Fig. 2-1D). Similarly, the peak for oxygen is higher than that for carbon. The pRDF for hydrogen atoms bonded to oxygen has the highest peak, followed by the pRDF for hydrogen atoms bonded to nitrogen atoms, followed by the pRDF for oxygen atoms. The pRDF for oxygen exhibits a small peak at 1.8 Å followed by a larger peak at 2.7 Å. The first peak is due to solvent hydrogen atoms, while the second is due to solvent oxygen atoms. The largest peak for the oxygen pRDFs lies at a greater distance than the first peaks of the pRDFs for hydrogen attached either to oxygen or nitrogen because oxygen has a larger radius than hydrogen and solute oxygen atoms generally are separated by a hydrogen atom.

Protein atoms are grouped using both coarse and fine definitions in order to test the degree of specification required to accurately reconstruct the hydration layer. The coarser specification groups heavy atoms by element (C, N, O, and S), with hydrogen atoms grouped according to their bonded heavy atom (CH, NH, OH and SH). A more refined definition distinguishes between every atom type for each amino acid separately (~300 types, Table 2-1). The utility of the finer definition is evidenced by the fact that pRDFs for O in the coarser definition differ from that for the O\textsuperscript{2-} of aspartic acid (Fig. 2-3A). The difference in accuracy of the hydration layer reconstruction using the two definitions is quantified below.
Figure 2-3. Specificity and transferability of pRDFs. A) Comparison between coarser and finer definitions of atom types. B) Comparison between pRDF’s calculated from Ub, Mb and HEWL.

The similarity of the pRDFs for the 3 proteins indicates that there exists a universal set of pRDFs applicable to a wide variety of proteins.

The pRDFs calculated from MD simulations for the three proteins are very similar (Fig. 2-3B). This identity suggests that pRDFs evaluated for one protein can be used to predict the solvent density around other proteins. This possibility is tested in the next section where the pRDFs determined from one protein are used to predict the hydration shell density simulated for another protein.
2.4.2 Reconstructing the hydration shell: HyPred

This HyPred reconstruction method using the coarser definition of atom types displays fewer features than obtained using the more refined definition of atom types, and generally the deviations from the average density are reduced (Figs. 2-1,2-4). The use of the finer definition of pRDFs produces a very similar solvation pattern as the explicit solvent simulations for each of the three proteins. Considering Ub, for example, a region of very high density (red in Fig. 2-1B) is present near the hydrogen atoms bonded to the nitrogen atom of the Arg54 side chain in both the reconstruction and the MD simulation. Just beyond that high density region is a regime of very low density that also appears in both the reconstruction and the MD simulation. Generally, very high density regions tend to be adjacent to regions with very low density.

Although most features are well reconstructed, some discrepancies exist. Four high density regions near Glu24 (green) are present in the MD simulations, but three of these become smeared into a single high density region in the reconstruction. Another high density region near Gln40 (magenta) appears in both the reconstruction and the MD simulation, although the region is much denser in the simulation, and its accompanying depletion layer has much lower density. Two of the three high density regions between Ala46 (orange) and His68 (blue) are well resolved in the reconstruction. The high density near Ala is associated with the amide hydrogen of the backbone nitrogen, demonstrating the importance of performing the reconstruction at the level of individual atoms types. All Ub reconstructions overestimate the density of the depletion layer separating the first and second hydration layers between Ala46 and Arg5
Figure 2-4. **HyPred reconstructions.** Solvent density around **A)** Mb and **B)** HEWL. For each protein, the solvation layer is obtained from either MD simulations or the pRDF derived from the simulations for that protein, and for the average of the other two proteins to test for transferability (lower right panel in A) and B)). Only one layer of cubes is shown but protein atoms within a 3 Å of the layer are shown. Thus some proteins atoms can be seen above the slab.
Excellent agreement also is obtained for Mb and HEWL using the finer atom type definition for the pRDFs (Fig. 2-4). Both the reconstructed and simulated hydration layers for Mb contain a region of very high density followed by a region of very low density near the hydrogen atoms bonded to the nitrogen atoms of Arg31 (red). Regions of high density similarly appear near the hydrogen atoms bonded to nitrogen atoms of Lys96 (green), near the hydroxyl hydrogen of Thr95 (blue), and near Glu148 (magenta) in both the reconstruction and the MD simulation.

For HEWL, a region of high density near Lys1 (red) is present in both the reconstruction and MD simulation. Another high density region near the carbonyl oxygen of Phe34 (green) is evident in the reconstruction and MD simulation, while two high density regions followed by low density regions near Lys13 (blue) are found in both. Numerous high density regions between Lys13 and Val120 appear in the MD simulations but the regions become smeared out in the reconstruction.

As further demonstration of the importance of calculating the reconstruction using data from MD simulations in which the protein atoms remain stationary, we compare this immobile protein reconstruction to one constructed from simulations in which all the atoms are mobile (Fig. 2-2). The time-average solvent density surrounding Ub is much more depleted near the protein when Ub is permitted to move during the simulations (Figs. 2-2A,B). The extent of
protein motion and the accompanying change in solvation are highlighted by two protein
snapshots (Fig. 2-2C,D).

2.4.3 Transferability.

The technique presented in this paper would be of limited predictive value if it could not
enable accurate prediction of the hydration of proteins from the pRDFs determined for other
proteins. To check for transferability, the pRDFs obtained from Ub are used to reconstruct the
hydration shells around Mb and HEWL. The reconstructions are quite similar to those obtained
using their own pRDFs (Fig. 2-3).

Table 2-2: Accuracy of the density reconstruction at 0.5 Å grid spacing

<table>
<thead>
<tr>
<th>Protein</th>
<th>pRDF&lt;sup&gt;b&lt;/sup&gt;</th>
<th>R</th>
<th>RMSD</th>
<th>R*</th>
<th>pRDF&lt;sup&gt;b&lt;/sup&gt;</th>
<th>R</th>
<th>RMSD</th>
<th>R*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ub</td>
<td>Ub</td>
<td>0.13 (0.14)</td>
<td>0.93 (0.98)</td>
<td>0.43 (0.46)</td>
<td>Mb,HEWL</td>
<td>0.14</td>
<td>0.99</td>
<td>0.46</td>
</tr>
<tr>
<td>Mb</td>
<td>Mb</td>
<td>0.16 (0.17)</td>
<td>1.13 (1.17)</td>
<td>0.49 (0.52)</td>
<td>Ub,HEWL</td>
<td>0.17</td>
<td>1.18</td>
<td>0.51</td>
</tr>
<tr>
<td>HEWL</td>
<td>HEWL</td>
<td>0.17 (0.18)</td>
<td>1.10 (1.15)</td>
<td>0.51 (0.54)</td>
<td>Ub,Mb</td>
<td>0.18</td>
<td>1.17</td>
<td>0.53</td>
</tr>
</tbody>
</table>

<sup>a</sup>Values in parenthesis are for the coarser definition of atom types (C, N, O, S, CH, NH, OH, SH)

<sup>b</sup>Column indicates the protein whose pRDF is used in reconstruction.

To quantify this agreement between the reconstruction and the MD simulations and any
differences resulting from changes in the protocol, three metrics are used (Table 2-2). The first
is the real space R factor (46),
where \( \rho_{o,i} \) is the average solvent density in cube \( i \) as calculated from the MD simulation, \( \rho_i \) is the reconstructed density for that cube, and the summation runs over cubes that lie within 8 Å of the protein. The second error measure is the RMSD between the two densities (46),

\[
RMSD = \sqrt{\frac{N \sum_i (\rho_{o,i} - \rho_i)^2}{\sum_i \rho_{o,i}^2}}
\]

The RMSD weighs more heavily the presence of regions with large disparity between the reconstruction and the MD simulation than the real space R factor. Because the R factor and RMSD strongly depend upon the extent of the bulk solution that is included in the calculation, we introduce a new measure that is not as strongly dependent upon the amount of bulk solvent in the reconstruction, provided that the simulation is long enough that bulk solvent density fluctuations, i.e., “noise”, is low. This third measure \( R^* \) is defined as

\[
R^* = \frac{\sum_i |\rho_{o,i} - \rho_i|}{\sum_i |\rho_{o,i} + \rho_i - 2\rho_s|}
\]

where \( \rho_s \) is the bulk solvent density. Bulk solvent should not affect \( R^* \) significantly because far from the protein, both \( \rho_{o,i} \) and \( \rho_i \) should equal \( \rho_s \), and the contribution of the bulk solvent to the numerator and denominator should each vanish.
All three metrics confirm that the reconstructions for each of the proteins using the average of the pRDFs of the other two proteins, agrees well with the reconstruction using the proteins own pRDFs (Table 2-2). When the pRDFs derived from the MD simulations of HEWL and Mb are averaged together and used to predict the hydration shell of Ub, the reconstruction is only marginally worse than the original obtained using Ub’s own pRDFs (R = 0.14 versus 0.13). These results demonstrate the transferability of the pRDFs to predict the hydration shell of other proteins.

2.4.4 Other factors influencing accuracy

We investigate which other features of our model produce improvements over the methodology of Pettitt et al. other than the use of a fixed protein conformation in the MD simulations. These new features include both a finer grid spacing (0.5 versus 2 Å) and atom type definitions (3 versus ~300+), and our separation distance is defined between the cube center and the scaled van der Waals surface rather than to the atom nucleus. The finer grid spacing permits a higher resolution reconstruction. But, the higher resolution has multiple features that degrade the performance as defined by the numerical measures. The 0.5 Å spacing is smaller than the peak widths of the pRDFs. This variation becomes averaged out when using 2 Å cubes, and the predicted map is inferior (Figure 2-5). In contrast, the averaging across the 64-fold larger cubes reduces the level of statistical noise. As a result of the smoothing and the increased statistical accuracy, the R factor and RMSD, for Mb, decreases from 0.16→0.042 and 1.13→0.22, respectively. The use of larger cubes impacts more on the RMSD because this
metric is more sensitive to larger discrepancies. The improved R factor and RMSD of the coarse grained reconstruction do not imply that the reconstruction is better at a lower resolution; rather, predicting the hydration layer at lower resolution is an easier task.

The improvement in the reconstruction using the finer atom type definition is evident in all three metrics (Table 2-2). Although the coarser model yields worse results, the data from this model contain reduced noise and thus can be used when insufficient data are available for constructing the pRDF of a specific atom type within some distance range when an unnatural amino acid is present, when an amino acid contains modifications, or for systems other than proteins.

When the separation $r$ in the pRDFs is defined following Pettitt as the distance to the nearest nucleus rather than to the nearest van der Waals surface, the R factor of the reconstruction slightly increases from 0.13 to 0.15 for Ub, 0.16 to 0.18 for Mb, and 0.17 to 0.18 for HEWL. Additional sources of errors may arise from ignoring influences from atoms other than the closest atom and their orientations (e.g., a hydrogen bond has a favored orientation that imparts angular and distance correlations to the local density dependence), from being on a convex versus a concave surface, from the secondary structure of the nearby atoms, and from nearby hydrogen bond donor/acceptors being internally satisfied versus not being internally satisfied. Despite ignoring all of these details, the HyPred method faithfully reproduces the MD simulations.
2.4.5 Comparison to bound waters in crystal structures

Some attempts at predicting the locations bound waters have employed information from X-ray crystal structures where highly ordered solvent sites within or near proteins can be identified (12-14). Others have attempted to predict crystallographic waters directly from MD simulations (47, 48) and from predictions of hydration shell densities (49). However, only a small number of highly ordered solvent sites can be determined for each protein, and inconsistencies sometimes arise between the sites assigned for different crystal structures of the same protein (50).

Here we test HyPred’s ability to predict the positions of high occupancy water molecules in experimental crystal structures. This comparison assumes that crystallographic waters are located at regions of high solvent density in the simulations. A multitude of factors argue against a one-to-one correspondence, (36, 51) including constraints imposed by crystallographic contacts, differences in temperature and buffer conditions between the simulation and the crystal structures, some assigned crystallographic waters might instead be ions or other solvent molecules, series termination errors may produce ripples that are misidentified as waters, and an excess water molecules may be used to over fit crystallographic data. For example, low temperatures can increase the occupancy in the crystal structures while the solvent density near charged amino acids is altered by the presence of solvated ions. A study of 10 T4 lysozyme crystal structures finds that 62% of the 20 most frequently occupied sites are conserved (50). Thus, it is unlikely that any comparison with crystal structures and MD
simulations would predict more than 60% of crystallographic water molecules unless the MD simulation is performed under the same conditions as the experiment. Additionally, a comparison is inherently limited by the accuracy of the underlying MD simulations, which are susceptible to systematic errors due to inaccuracies in the TIP3P water model and the force field.

Nevertheless, we compare water molecules observed in crystallographic structures to regions of high solvent density in both the MD simulations and the HyPred reconstructions. Water molecules are predicted at positions where the solvent density in a $(0.5 \text{ Å})^3$ cube is above a threshold level except when another cube with a higher density lies within 2.8 Å (the diameter of a water molecule). Two tests are performed to assess the accuracy.

The first test involves calculating the percentage of the correctly predicted crystallographic water molecules as a function of the total number of predicted water molecules. The number of predicted molecules is varied by adjusting the threshold level. For Ub, Mb and HEWL, the MD simulations are able to predict between 1/4-1/2 of the crystallographic water molecules, depending on the protein (Fig. 2-6). The HyPred results are slightly worse, but still much better than the control where water molecules are randomly placed within 3 Å of the protein’s surface. The accuracy, defined as (true positives)/(total predicted), generally is higher when a higher threshold is used and fewer positions are predicted (Fig. 2-6b).
Figure 2-5: Hydration density at 2 Å resolution. The hydration shell reconstruction of Mb shows far fewer features at a resolution of 2 Å than at 0.5 Å.
Figure 2-6: Predicted water molecules around proteins. For MD (black trace), HyPred (blue trace) and randomly placed waters within 3 Å of the protein surface (red), the percentage of crystallographic water molecules with a predicted water molecule within 1 Å is calculated as function of the number of predicted water molecules as the density threshold is progressively decreased (bounded at $r < 0.5$ protons/Å$^3$). There are 58, 74 and 101 water molecules in the Ub, Mb and HEWL crystal structures, respectively (vertical black lines).
In a second test, we calculate the fraction of crystallographic water molecules with a predicted water molecule within a cutoff distance (Fig. 2-7). The density threshold value is adjusted for each protein so that the same number of water molecules is predicted as observed in its crystallographic structure. When this procedure is performed using the solvent density from the Ub MD simulation, 17 of the 58 waters identified in the crystal structure have a predicted water molecule lying within 1 Å. When the same procedure is performed on the HyPred reconstructed density, 10 of the crystallographic waters have a predicted high occupancy site within 1 Å. For the 74 crystallographic water molecules in Mb, the MD simulations and HyPred reconstruction correctly predict 18 and 10 molecules within 1 Å, respectively. Of the 101 crystallographic water molecules in HEWL, the MD simulation and HyPred reconstruction correctly predicts 13 and 7 molecules within 1 Å, respectively. For the three proteins, none of the crystallographic waters are nearby the appropriate number of randomly placed water molecules.

In spite of the aforementioned caveats concerning the ability of MD simulations to reproduce crystallographic water molecules, these two tests provide experimental support of the validity and utility of the HyPred reconstruction method.
Figure 2-7: Water molecules around proteins. A) Accuracy of predicted crystallographic water molecules from MD simulations and HyPred are compared to random distributions. The Y-axis is the fraction of crystallographic water molecules with a predicted water molecule within a cutoff as a function of the cutoff. B) Crystallographic water molecules (red) are compared to predicted locations from the MD simulation (blue).
2.5 Conclusions

We present a significantly improved model for rapidly calculating the solvent density around a protein. The reconstructions assuming the interactions between the water and protein are well represented by pRDFs for the closest protein atom obtained from MD simulations in which only the water is allowed to move. Radial distributions are found to be independent of protein, thereby enabling the prediction of hydration layers for new proteins without the need for additional MD simulations. Our use of residue-dependent atom types improves the accuracy of the reconstruction of the hydration layer from the radial distribution functions by implicitly including the influence of nearby side chain atoms. Improvements might accrue by explicitly including the influence of the second nearest neighbor and other factors, including the local curvature of the protein’s van der Waals surface, the angle formed between the vector connecting the cube and the protein atom with the bond vector connecting the atom and the second nearest neighbor, and using multi-body correlation functions. Using the TIP4P or TIP5P water model in the MD simulations might improve agreement with experiment. Future applications include predicting the hydration layer surrounding RNA, DNA, and biological membranes. It might be possible to use a similar method as presented here to predict residence times of water molecules, and the preferred orientations of water molecules around proteins. It might be possible to estimate the free energy of solvation of proteins, with some modifications, using a similar method as Lazaridis(52). McLachlan has calculated solvation free energies based on accessible surface areas(53). It might be possible to calculate the contribution of each atom type to the free energy of solvation and compare to solvation
parameters of McLachlan. A website for doing HyPred calculations can be found at http://godzilla.uchicago.edu/cgi-bin/jouko/HyPred.cgi.
3. Modeling the hydration layer around proteins:

Applications to small and wide angle X-ray scattering

3.1 Abstract

The contents of this chapter have been published (54). Small/wide angle X-ray scattering (SWAXS) experiments can aid in determining the structures of proteins and protein complexes, but success requires the accurate computational treatment of solvation. We compare two methods to calculate SWAXS patterns. The first approach uses all-atom explicit solvent molecular dynamic simulations. The second, far less computationally expensive method involves prediction of the hydration density around a protein using our new HyPred solvation model that is applied without the need for additional MD simulations. The SWAXS patterns obtained from the HyPred model compare well to both experimental data and to the patterns predicted by the MD simulations. Both approaches exhibit advantages over existing methods for analyzing SWAXS data. The close correspondence between calculated and observed SWAXS patterns provides strong experimental support for the description of hydration implicit in the HyPred model.

3.2 Introduction

SWAXS data provides information on the size and shape of biological molecules in solution. The considerable resurgence in SWAXS experiments during the last decade (55, 56) emerges, in part, from the availability of high flux synchrotron sources whose use reduces
sample requirements and measurement times, enabling structural studies under a variety of solvent conditions. In addition, the real space reconstruction methods of Svergun and coworkers for producing 3D envelopes (18, 57) from SWAXS data greatly enhance the utility of SWAXS for characterizing the structure of proteins and complexes. The combination of this improved utility and further theoretical advances greatly enhances the power of SWAXS for characterizing motions that are incompatible with a crystal lattice, for studying proteins that are too difficult to crystallize and/or too large for NMR methods (58), and for analyzing systems under a wide range of solution conditions not compatible with crystallography. Additional theoretical advances are of interest to facilitate extraction of information on protein secondary structure and protein folds from the wider angle portion of the SWAXS pattern (59).

The utility of SWAXS for testing structural models of proteins depends on the ability to accurately calculate the SWAXS pattern corresponding to a given protein structure. The greatest challenge in these calculations involves describing the influence of the solvent to enable extracting the protein’s contribution from the total scattering pattern (33, 60, 61). This extraction is required because SWAXS patterns depend upon the distances between all pairs of electrons in the system, including distances involving the electrons in the solvent. The solvent contributes to the SWAXS pattern in two ways. The first contribution from solvation arises from the scattering due to the bulk solvent and the excluded volume arising from the displacement of solvent by the protein, while the second emerges because of the variable solvent density in the hydration shell layer surrounding the protein. It has been previously found that the average density of the hydration layer deviates from that of bulk solvent (33, 62). Correcting for the
scattering by the bulk solvent is comparatively straightforward, while modeling the hydration shell layer and its influence on the SWAXS pattern remains a challenge.

The hydration shell is usually described by adding a layer of uniform (excess) electron density around the surface of the protein as implemented in the widely used CRYSOL program developed by Svergun et al. (63). CRYSOL is the current standard for SAXS calculations and has played an important role in the resurgence of SAXS. In actuality, however, the hydration shell is nonuniform, with successive layers of positive and negative deviations from the bulk density. Furthermore, the solvent density exhibits Å scale variations due to the different chemical properties of the atoms on the surface of the protein. Despite this variability, SWAXS data are generally analyzed by optimizing the fit of a uniform hydration shell density to experimental scattering data (63). Another method for calculating SWAXS patterns employs coarse-grained water molecules, whose positions are obtained by overlaying the protein on an equilibrated water box and removing the waters that are not in the hydration shell and by assigning weights to the water molecules to optimize agreement with experiment (64). Bax et al. use a similar approach without coarse-graining (65).

A more realistic model of the hydration shell is required to enable improved calculation of SWAXS patterns and obviate the need to optimize adjustable parameters. Because MD simulations provide atomistic resolution for the solvent density around proteins without the need for adjustable parameters, Park et al. (66) have calculated SWAXS patterns from explicit atom MD simulations of water positions. The large expense of MD calculations implies that the routine use of SWAXS data for protein structure determination requires the development of
fast and accurate methods for calculating SWAXS patterns without the need for expensive MD simulations. Here we devise such an approach based on HyPred (17), a fast analytical method with atomic level precision for predicting the nonuniform solvent density near the surface of a protein. The SWAXS patterns obtained using HyPred’s treatment of solvation are demonstrated to agree well with experiment and those calculated from MD simulations, a comparison devoid of systematic errors in force fields and experiments. In addition, an analysis is provided for the influence of protein dynamics on scattering patterns.

3.3 Methods

3.3.1 Experimental SWAXS data

All data are collected as described previously (67). SWAXS data have been collected for multiple concentrations of Ub, HEWL, and Mb. The SWAXS patterns are linearly extrapolated to zero concentration. For Mb only, data from the three most dilute solutions (11.04 mg/ml, 8.83 mg/ml, and 4.52 mg/ml) are extrapolated to zero concentration for $q < 0.2 \text{ Å}^{-1}$ because the scattering intensity varies non-linearly with protein concentration at high concentrations. Data collected at high concentrations is used without modification for $q > 0.2 \text{ Å}^{-1}$ because the experimental error is smaller at higher concentrations and the scattering at wide angles is independent of concentration (68).

3.3.2 Molecular dynamics simulations

All-atom explicit solvent MD simulations have been performed at 4 °C for 50 ns for ubiquitin, (Ub, 1UBQ (39)), hen egg white lysozyme (HEWL, 6LYZ (40)), and myoglobin (Mb,
1WLA (41)), employing NAMD (42) and the CHARMM 27 all-atom force field (43). Counterions are added to compensate for the charges on the proteins and to match experimental conditions (67). A simulation of a buffer solution is also performed. Simulations for twenty additional proteins (1DF4, 1DXG, 1F94, 1HYP, 1L2P, 1TIF, 1UOY, 1US0, 1VCC, 1YZM, 2DOB, 2ZQE, 3G19, 3HGL, 3LE4, and four models from 2KN5) employ simulation boxes whose dimensions extend 15 Å beyond the proteins on all sides, except for the 1DF4, 1L2P, 1TIF, and 3G19 simulations where the dimensions extend 25 Å beyond the protein on all sides. Charges are neutralized with either Na\(^+\) or Cl\(^-\) ions, and protonation is set by assuming standard states for pH = 7. The density of bulk solvent is determined in each case from all water molecules farther than 10 Å from the protein. If the average computed electron density is less than 0.333 e\(^-\) Å\(^{-3}\) or greater than 0.335 e\(^-\) Å\(^{-3}\), the simulation box size is adjusted to give the correct bulk density. Only protein monomers (containing between 36 and 314 residues) are considered. In order to reduce computational expense, we have mostly chosen small proteins for study, but our methods can be applied to proteins of arbitrary size.

Energy minimization and equilibration proceed in several stages. The solvent and protein hydrogen atoms are first energy minimized for 2000 steps; then, with all atoms in the protein other than the hydrogen atoms immobilized, the temperature of each system is initially set at 1200 K, the systems are cooled to 4 °C over a period of 100 ps, and the systems are then equilibrated for 100 ps at 4 °C. This annealing enables the ions to move from their original positions. Then, the hydrogen atoms in the protein are fixed in place, and the systems are equilibrated for another 4500 ps. One simulation for Mb (and another for cytochrome c (1CRC))
is performed with the temperature initially set at 10 K and then raised by 10 K every 100 ps until reaching the final temperature of 4 °C. All protein atoms remain immobile throughout the course of the subsequent MD simulation to keep the structure from deviating from the crystal structure and to eliminate the smearing of the calculated SWAXS patterns due to the dynamics of the protein (as described below). One additional 50 ns simulation (after a 4.5 ns equilibration) of Ub is performed in which the protein atoms are allowed to move in order to analyze the role of protein thermal fluctuations. Electrostatic interactions are computed with particle mesh Ewald summations. A 1 fs time step is used, and snapshots are saved every 1 ps. The simulations enforce NVT (constant number of atoms, volume, and temperature) conditions.

3.3.3 Calculation of WAXS patterns from MD simulations

The Debye formula can be used to calculate the SWAXS pattern of a solution from the atomic coordinates of all species in the system,

\[
I(q) = \sum_{i,j=1}^{N} f_i(q) f_j(q) \sin \frac{q r_{ij}}{q r_{ij}}
\]

where \( q = \frac{4\pi \sin \theta}{\lambda} \) is the momentum transfer, \( \lambda \) is the wavelength of the X-rays, \( 2\theta \) is the scattering angle, \( r_{ij} \) is the distance between atoms \( i \) and \( j \), \( N \) is the number of scattering particles, and \( f_i(q) \) is the scattering factor for atom \( i \). Because the MD simulations involve systems of finite size, artifacts may arise at small \( q \) where the calculated scattering may appear similar to the scattering from an object with the shape of the simulation box. This artifact can be reduced by adding a uniform electron density equal to that of the bulk solvent outside of the
finite system box, thereby effectively making the system size infinite. This addition of the outside density is implemented using Babinet's principle (70), which states that the scattering from a uniform object is the same as the scattering from a hole of the same shape. Hence, adding density everywhere outside of the system box is equivalent to subtracting electron density from every point inside the system box. Babinet's principle is expressed with,

$$f_{\text{outside}}(q) = -f_{\text{inside}}(q) + \delta(q)$$  \hspace{1cm} (3)$$

where $$f_{\text{inside}}(q)$$ is the scattering factor for an object of uniform density that is enclosed within a boundary, and $$f_{\text{outside}}(q)$$ applies for an object with uniform density everywhere outside of the boundary. The delta function can be ignored since scattering is not measured at $$q=0$$. The subtraction of electron density from each point in the system is achieved using a single dummy atom with the same shape and size as the system and with an electron density equal to the negative of the bulk electron density. The Debye formula requires that all particles in the system be spherically symmetric. Hence, the scattering from a dummy atom shaped like a rectangular prism cannot be calculated using this formula because the rectangular box lacks spherical symmetry. However, a spherically symmetric system can be constructed by discarding atoms outside of a large sphere centered about the protein's center of mass. After extracting this spherical system, a uniform bulk solvent electron density is added everywhere outside of the system sphere, and the Debye formula (Eq. 3-1) is used. This approach, however, produces an excess electron density near the outside surface of the sphere because atoms centered inside but extending outside the boundary of the sphere also contribute electron density outside. Similarly, the density just inside the sphere is reduced because of the removal outside
of the sphere of explicit atoms that are centered outside but that would contribute to the
density inside the sphere. Therefore, the density added to produce an average uniform electron
density at large distances from the solvent cannot be introduced with a sharp boundary, but
must instead begin increasing a few Ås before the inside boundary of the sphere and only attain
bulk density a few Ås outside of it. The appropriate density can be calculated by convoluting the
electron density of the sphere with the electron density of the two hydrogen atoms and one
oxygen atom from one water molecule. The electron density due to the solvent at any
particular point \( r \) can be found by integrating over all space the probability of finding a solvent
atom at \( r_2 \) multiplied by the contribution of a solvent atom at \( r_2 \) to the electron density at \( r \). This
density is just the convolution of a sphere with the density of two hydrogen atoms and an
oxygen atom. Using the convolution theorem, which states that
\[
F[f \otimes g] = F[f]F[g]
\]  \( 4. \)
where \( F[f] \) and \( F[g] \) are the Fourier transforms of \( f \) and \( g \), respectively, and \( f \otimes g \) denotes the
convolution of \( f \) and \( g \), the scattering factor of the convoluted sphere emerges as equal to the
scattering factor of the sphere multiplied by the sum of the scattering factors of two hydrogen
atoms and an oxygen atom,
\[
f(q) = -f_{\text{sphere}}(q) \left[ \frac{2f_H(q) + f_O(q)}{2f_H(0) + f_O(0)} \right]
\]  \( 5. \)
where \( f_{\text{sphere}}(q) \) is the scattering factor of a hard sphere, and \( f_H(q) \) and \( f_O(q) \) are the scattering
factors of a hydrogen and an oxygen atom, respectively. Eq 4 is used as one of the scattering
factors in eq 3-1. Now that spherical symmetry has been restored, the Debye formula is applied
to each snapshot of a trajectory in order to calculate the SWAXS patterns from the MD simulation every 10 ps, with subsequent averaging. Some residual error remains when the density is made uniform outside of the sphere due to loss of correlation in electron density between points lying just inside and just outside of the sphere. If this error is significant, it should depend upon the sphere size. A sphere of radius 39 Å is used for Ub, HEWL, and Mb. The sphere sizes for the other proteins are set such that no protein atom is closer than 8 Å to the boundary of the sphere. Tests of spheres with radii of 49 Å, for Ub, HEWL, and Mb yield very similar scattering patterns to those for the smaller spheres for each of the three proteins. This implies that losing information about density correlation between points inside and outside of the sphere imparts minor errors. The calculated SWAXS patterns converge rapidly: those from the first 9 ns and the full 50 ns of the MD simulations after equilibration are very similar for Ub, HEWL, and Mb (Figure 3-1). The differences between the calculated SWAXS patterns of the two simulations of Mb with different equilibration procedures are likewise minor (Figure 3-2).
Figure 3-1: Evidence of sufficient sampling. The SWAXS patterns are almost identical when calculated from the first 9, 25 and 50 ns of the MD simulation after equilibration.
Figure 3-2: Equilibration. The black curve is from the MD simulation of Mb in which the system temperature is initially set to 1200 K and then lowered to 4 °C, and the red curve is from the MD simulation of Mb in which the system temperature is initially set to 10 K and then increased to 4 °C.
A simulation of the buffer solution enables subtraction of the background buffer contribution from the composite intensity obtained from the simulation with both protein and buffer. Finite-size artifacts largely cancel when the SWAXS pattern of the buffer solution is subtracted from the SWAXS pattern of the protein in solution. The scattering from the buffer solution is subtracted from the scattering from the protein in water according to the equation,

\[ I(q) = I_{\text{protein}}(q) - \mu I_{\text{buffer}}(q) \]

where \(\mu\) is the ratio of the average number of solvent electrons in a frame of the protein simulation relative to that for the buffer alone. Computations with the Debye formula are accelerated by calculating histograms of interparticle separations for each pair of atom types using a bin size of 0.001 Å. This small bin size is necessary because the wide angle scattering is dominated by the scattering from the solution for both the protein and buffer systems. A small error in the scattering in either the protein or buffer simulations produces a large error in the calculation of wide angle scattering from the protein.

### 3.3.4 Calculating SWAXS patterns using the cube method

Another way of describing the influence of solvent involves discretizing the electron density into \((0.5 \text{ Å})^3\) cubes and specifying the electron density within each cube either from the average of snapshots taken from the MD trajectory at 1 ps intervals or from the HyPred method. These computed SWAXS patterns utilize the averaged solvent densities and the Cartesian coordinates of the immobile atoms in the protein. The coarse graining of the electron density into cubes eliminates information concerning correlations in density between points in...
the hydration shell. Cubes are ignored when they are situated in the solution further than 8 Å from any of the nuclei of the atoms on the protein's surface. Babinet's principle (70) implies that contributions from the bulk solvent can be described by subtracting the average solution density from each of the cubes. Because the cubes are not spherically symmetric, the Debye formula cannot be applied, and instead, the scattering is computed in three-dimensional reciprocal space and then spherically averaged numerically. More specifically, the SWAXS pattern is obtained by averaging over 200 spherically distributed scattering vectors, \( \mathbf{q} \) (i.e., averaging over 200 orientations), for each magnitude of \( q \),

\[
I(q) = \langle I(q) \rangle = \left\langle \sum_{n=1}^{N} f_n(q)e^{i\mathbf{q}\mathbf{r}_n} \sum_{m=1}^{N} f^*_m(q)e^{-i\mathbf{q}\mathbf{r}_m} \right\rangle \tag{7}
\]

where \( N \) is the total number of atoms and cubes used in the calculation, and \( \mathbf{r}_n \) is the position of particle \( n \). The orientations of the scattering vectors \( \mathbf{q} \) are chosen with a uniform distribution for the azimuthal angle of the \( \mathbf{q} \) vectors and for the cosine of the polar angle. Only scattering vectors in one hemisphere are needed because the scattering for \( \pm \mathbf{q} \) are identical.

The scattering factor for a single cube is given by

\[
f(q) = 8\rho \frac{\sin(q_xa/2)\sin(q_ya/2)\sin(q_za/2)}{q_xq_yq_z} \tag{8}
\]

where \( a \) is the length of an edge of the cube, \( \rho \) is the electron density in the cube, and Eq. (7) is used in Eq. (6) for each solvent containing cube. The cubes are treated differently from the
atoms of the protein. Since the cubes are evenly spaced, the contribution to Eq. (6) from the cubes is determined by fast Fourier transformation, while the contribution from the protein is evaluated by direct Fourier transformation (Method 2: cube method) because the atoms of the protein are not evenly spaced.

3.3.5 HyPred method

HyPred proceeds by calculating a set of proximal radial distribution functions (pRDF), defined below, for ~ 300 different categories of atoms in proteins using all-atom, explicit solvent MD simulations (17). This set of ~300 atom types distinguishes between atom and amino acid type. The pRDFs describe the hydration layer near the protein’s surface atoms and are used to predict the solvent density in a discrete description. The pRDFs are calculated by partitioning the simulation box into (0.5 Å)$^3$ cubes which are assigned to the solute atom whose (scaled) van der Waals surface is the closest, and the separation is stored. The densities and separations of the cubes assigned to each particular atom type are then used to calculate the entire pRDF for that atom type. The hydration shell is recreated using the same grid by inverting the process used to generate the pRDFs. We store the separation between each cube outside of the protein and the solute atom with the nearest scaled van der Waals surface. The solvent density in each cube is assigned the value of the pRDF associated with that solute atom type and the separation. The model and the distribution functions for a given protein have been validated by their ability to reproduce the hydration layer from the MD simulations for that protein as well as those for other proteins. Cross sections of the average hydration shell
electron densities obtained from MD simulations and HyPred predictions are illustrated in Fig. 3-3 (insets) and Fig. 3-4.

Figure 3-3: Comparison between calculated and experimental SWAXS patterns. A) HEWL and scaled buffer solution scattering as calculated from MD simulations. B-D) Comparison of calculated background subtracted SWAXS pattern of Ub, HEWL and Mb with experiment. The protein is held immobile. Traces are scaled to set I(q=0)=1 except for Mb where the three curves are scaled to match at q=0.64 Å⁻¹. A cross-section of the MD(TIP3P) and HyPred model of the hydrated protein (insets, with legend).
Figure 3-4: Cross section of the hydration shell electron densities of Ub, HEWL, and Mb as determined by MD simulations and predicted by HyPred. High density is represented by red, low density is represented by blue, and density approximately equal to bulk is represented by white.
When HyPred is used to calculate the SWAXS pattern of a protein, simulation data for that particular protein are omitted, so the pRDF is obtained as an average of pRDFs constructed from data for the other eighteen proteins considered. Since the appearance of the publication describing HyPred, several improvements have been made to HyPred. The present work employs an enhanced procedure for calculating the pRDFs beyond those obtained from the original HyPred method (17). The first enhancement corrects for hydrogen bonding between protein and solvent by adding a specification of the dependence of the pRDFs on the angle between the vector connecting the center of the cube and the protein atom that is closest to the cube and the bond connecting the atom to its nearest bonded atom (Fig 3-5). The bin size is $15^\circ$, and cubes are also categorized according to the atom type of the second closest atom in space. In this case the atom type is specified by elemental type of the second nearest heavy atom (e.g., S, O, N, or C) or, if the second nearest neighbor is a hydrogen atom, the specification is according to the heavy atom to which the H atom is bonded. Cubes are also categorized according to whether they are near a convex or concave region of the protein surface. Finally, the number of reference proteins used to generate the pRDFs has been increased from 3 to 19 (listed above).
3.4 Results and Discussion

3.4.1 Simulations and comparison to experimental data

SWAXS patterns are calculated using 5000 snapshots taken from 50 ns MD simulations for each of three globular proteins, hen egg white lysozyme (HEWL), ubiquitin (Ub), and myoglobin (Mb). The protein is held immobile in these simulations, and only the solvent molecules are allowed to move (Fig. 3-3). The scattering from the composite protein plus solvent is peaked at q=0 and also contains a broad peak at 2.1 Å⁻¹ that appears in the scattering from the solvent alone and arises from density variations of the water molecules. This peak is removed upon subtracting the buffer scattering after scaling the latter by the ratio of average number solvent electrons in the protein solution relative to the number in the buffer solution. The latter procedure is required because the average numbers of solvent electrons in frames of...
the Ub and buffer MD simulations are 79655 and 83152, respectively, so the buffer SWAXS scattering is scaled by a factor of 0.958 before subtracting the buffer SWAXS scattering to compensate for this difference in the contribution of the buffer to the SWAXS scattering.

The resulting buffer-subtracted scattering patterns for HEWL and Ub agree quite well with the experimental data over the entire range of $q$ (Fig. 3-3b,c), while some deviation appears for Mb particularly at small angles (Fig. 3-3d), perhaps due to concentration effects. SAXS patterns can vary with protein concentration because of interparticle interference as well as changes in internal electron density variations when the electron density of the solvent changes, and increasing dynamics of the protein when it has more room to undergo structural fluctuations in dilute solutions (68). These effects are large and nonlinear in Mb but are far smaller at wide than at small angles. This explains why the calculations reproduce the scattering at wide angles, but deviate at small angles. Figure 3-6 illustrates Figure 3-3d scaled such that the experimental data and the SWAXS pattern calculated from the MD simulation match at small angles.
Figure 3-6: Comparison between calculated and experimental SWAXS patterns. Comparison of calculated background subtracted SWAXS pattern of Mb with experiment. Traces are scaled to set $I(q=0)=1$.

Zinke et al. (68) have studied the SWAXS pattern of hemoglobin as a function of concentration and find large nonlinear changes at small angles but negligible changes at wide angles.

Makowski et al. (67) have also studied the dependence of SAXS on concentration for proteins, including Mb. Figure 3-7 displays several SWAXS patterns of Mb for various concentrations of
Mb, alongside the SWAXS pattern of Mb calculated from the MD simulation. The intensities at small angles vary greatly, and the SWAXS pattern calculated from the MD simulation overlaps more with the lower Mb concentration data sets, as expected given that the simulations are in the very low concentration limit. Other potential sources of error could arise from differences between the crystal and solution structures near the chromophore binding pocket.

Figure 3-7: The SWAXS patterns for concentrations of Mb ranging from 200 mg/mL to 4.52 mg/mL and the SWAXS pattern calculated from the MD simulation.
Figure 3-8: Comparison between calculated and experimental SWAXS patterns for cytochrome c. Comparison of SWAXS pattern calculated from an MD simulation and HyPred with the experimental data.

Figure 3-8 presents a comparison between the SWAXS pattern calculated from the MD simulation of cytochrome c and experiments. Deviations between the calculated SWAXS pattern and experiment also may arise from differences between the TIP3P water model and
real water molecules. Different water models might result in different average densities for the hydration shell and therefore impact the SWAXS patterns of proteins.

Because a protein’s structure varies dynamically, the experimental scattering may not be well represented by the single structure used in our MD simulations. To investigate the influence of protein motion, the SWAXS pattern of Ub is calculated from an all-atom MD simulation in which protein motions are permitted. The resulting pattern does not agree as well with experiment as the scattering from the static structure, perhaps indicating deficiencies of the force field. The intensity at q=0 is higher by 15% than that utilizing the MD simulation in which the protein is immobile (Fig. 3-9). This increase in intensity at q=0 is possibly due to a greater number of water molecules in the hydration shell of the dynamic protein than the number of water molecules in the simulation with the immobile protein. This difference arises from only eight extra water molecules on average in the hydration shell of the dynamic protein, which is an illustration of how sensitive SWAXS can be to hydration.
Figure 3-9: Influence of dynamics on scattering profiles for Ub. The calculated SWAXS patterns are divided by I(0) from the MD simulation with the protein held fixed, and the experimental data are scaled to unity at q = 0.

The resulting radius of gyration, $R_g = 12.80 \pm 0.10 \, \text{Å}$, is larger than both that from the simulation with the immobile protein, $12.62 \pm 0.07 \, \text{Å}$, and from experiment, $12.40 \pm 0.02 \, \text{Å}$, where the $R_g$ values and error bars are obtained from the second moment of $P(r)$, the electron pair distance distribution function as determined using the GNOM indirect Fourier transform procedure(71) and retaining scattering data over the range 0.06-0.6 $\, \text{Å}^{-1}$.
The experimental $R_g$ of HEWL is $14.61 \pm 0.04$ Å and departs from the experimental $R_g$ reported by Svergun, which is $15.4 \pm 0.2$ Å. Similar differences are seen with Mb. There are a number of potential reasons for this difference. $R_g$ may be a function of protein concentration, partly due to interparticle interference at higher concentrations. This effect is generally ignored because SAXS data taken at synchrotrons are usually collected at submillimolar protein concentrations. In addition, internal electron density variations alter $R_g$ when the electron density of the solvent changes. Moreover, a protein may undergo larger fluctuations in dilute rather than concentrated solutions, and this may lead to the inference of a greater $R_g$. Buffer conditions may also affect the $R_g$. The WAXS data from Ub, HEWL, and Mb are very well determined and have been collected multiple times.

SWAXS patterns have been calculated from MD simulations for a set of NMR structures of Mb (Figure 3-10). The SWAXS patterns calculated from the NMR structures agree better with the experimental data than the SWAXS pattern calculated from the crystal structure, suggesting that the NMR structures more accurately reflect the ensemble of solution structures than the crystal structure.
Figure 3-10: SWAXS patterns calculated from a crystal structure of Mb as well as from 4 different NMR structures of Mb. The number after the PDB code indicates the model number.

3.4.2 Calculations using average solvent density

We next examine whether the use of the average of the solvent density over the MD trajectory for each $(0.5 \, \text{Å})^3$ cube is adequate to produce accurate scattering patterns. Cube models have been used previously to calculate SWAXS scattering intensities (72-74), but our model provides a more accurate depiction of the hydration layer and hence, the averaged density.
approximation may provide accurate SWAXS patterns. The SWAXS patterns are compared using both methods for determining the solvent density within the cubes. The first approach uses the solvent density averaged over the 50 ns MD simulations, while the second approach employs the predicted densities from HyPred.

The scattering patterns for these two descriptions of the solvent are remarkably similar to the patterns obtained using the average of the scattering from the 5000 MD snapshots (Fig. 3). The slight discrepancy at $q > 1.3 \, \text{Å}^{-1}$ is not due to discretization errors arising from the use of $(0.5 \, \text{Å})^3$ cubes; discretization into either $(0.25 \, \text{Å})^3$ or $(0.75 \, \text{Å})^3$ cubes produces nearly identical patterns (Fig. 3-11). Hence, the difference at high $q$ mostly likely arises from the loss of density correlations induced by the approximations.

The SWAXS patterns calculated from the MD simulations and from HyPred agree quite well for all three proteins. The patterns for HEWL agree well until $q \approx 0.5-1 \, \text{Å}^{-1}$, while the agreement for Ub persists out to $2 \, \text{Å}^{-1}$. The agreement for Mb extends to $2 \, \text{Å}^{-1}$, except near $q = 0.2 \, \text{Å}^{-1}$. The SWAXS patterns for HEWL and Mb, obtained with either of the MD-based methods, are more similar to each other than the patterns calculated using the HyPred solvent model (Fig. 3-12). This indicates that further improvements in accuracy may be possible for some proteins. A comparison between the SWAXS pattern of cytochrome c calculated using HyPred, the SWAXS pattern calculated from the MD simulation, and the experimental data is displayed in Figure 3-9.
Figure 3-11: Effect of discretizing the solvent. Comparison between the SWAXS pattern of Ub as calculated from MD simulation with explicit solvent molecules and the pattern obtained using the averaged solvent density within cubes of edge length 0.75, 0.5, and 0.25 Å.
Figure 3-12: The influence of discretizing and averaging the electron density into (0.5 Å)$^3$ cubes. The protein is held immobile.

3.4.3 Comparison to other methods

Our I(q) calculations using explicit solvent simulations are very similar to those of Park et al. that have been reported for HEWL and Mb (66) (Fig 3-3 b,d). One important difference arises because Park et al. subtract the unscaled buffer scattering from the protein scattering, whereas
we scale the buffer scattering before subtraction by the ratio of the number of solvent electrons in the protein simulation to the number of solvent electrons in the buffer simulation (~90-99% depending on the protein and sphere size). Additionally, Park et al. use a 7 Å layer of water around the protein, while we place the protein in a sphere of water with constant electron density everywhere outside of the sphere. The treatment of Park et al. yields an alternative scaling for the buffer contribution which leads to the calculation of an 'excess' intensity that is not directly comparable to the background-subtracted intensities we report here for \( q \geq 2 \, \text{Å}^{-1} \) because of the different scaling of the scattering from the protein and buffer solutions (Park et al., 2009).

The discretization of space into cubes leaves no gaps in the solvent and improves the calculated SWAXS intensity at high \( q \), for example, as compared to CRY SOL, which uses spherical dummy atoms with negative electron density to subtract the contribution from the bulk solvent electron density. The benefit of the increased accuracy of discretization that emerges from using the cubes comes at the expense of a loss of spherical symmetry that invalidates the use the computationally rapid Debye formula. However, the solution scattering can be well represented by calculating the scattering pattern from the average of 200 different orientations. Using 450 orientations rather than 200 does not significantly alter the calculated SWAXS pattern. The validity of this method is confirmed in the previous section by comparing SWAXS patterns from the HyPred cube method to those from MD simulations. The total computation time for hydrating the protein and calculating the SWAXS pattern is ~30 minutes for Ub, ~40 minutes for HEWL, and ~45 minutes for Mb, on an Intel Pentium 4 2.8 GHz CPU. An
advantage of CRYSOL over the methods presented in this paper is that CRYSOL is faster by at least an order of magnitude, and for those who want to perform many SAXS calculations in a short amount of time, CRYSOL may be the better option. One deficiency with the HyPred model arises because it is based on MD simulations at low ionic conditions and therefore may not be able to accurately reproduce the hydration layer for other buffer condition. To maintain spherical symmetry, the CRYSOL program implements Babinet’s principle by introducing spherical dummy atoms, with negative electron density, which are placed directly on top of the real protein atoms. The use of spherical dummy atoms introduces spurious gaps with excess or deficient electron densities (63). In order to assess the improvement gained by using HyPred instead of a uniform hydration shell, such as the one applied by CRYSOL, the SWAXS patterns for the three proteins are also calculated using the cube method with a uniform hydration shell of thickness of 3 Å and an excess density of 0.03 e Å$^{-3}$, which is the same thickness and density as the hydration layer used by CRYSOL (using default hydration parameters). Figure 3-13 contrasts the SWAXS patterns calculated using HyPred and a uniform hydration shell. The impact of using a uniform hydration shell is substantial, and the SWAXS patterns calculated for Ub, HEWL, and Mb using a uniform hydration shell are very similar to the SWAXS patterns calculated using CRYSOL, suggesting that the majority of the difference between the HyPred cube method and CRYSOL is due to the difference in the treatment of the hydration shell rather than the other factors discussed (the treatment of excluded volume, the use of dummy atoms or discretization using cubes). In addition to calculating the SWAXS patterns of proteins with a uniform hydration shell, Figure 3-14 also presents calculated SWAXS patterns of proteins.
without a hydration shell and of proteins in vacuum. These calculated SWAXS patterns have a significantly higher intensity at \( q<0.2 \ \text{Å}^{-1} \) because of the greatly enhanced contrast between the electron density of the protein and vacuum (compared to solvent). Because the SWAXS patterns are scaled to unity at \( I(0) \), the patterns calculated using CRY SOL appear to have a significantly lower intensity at \( q>0.2 \ \text{Å}^{-1} \).

**Figure 3-13: Comparison between various hydration models and CRY SOL.** The protein is kept immobile.
Based on a comparison of the SWAXS patterns, our HyPred model provides a superior representation of the MD simulations than the model used in CRYSOL with the default parameters for the hydration shell density and atomic radii using 50 spherical harmonics (Fig. 4). The use of 50 spherical harmonics is necessary and sufficient to calculate SWAXS patterns to wide angles. The HyPred method produces significantly superior SWAXS patterns than CRYSOL for $q < 0.6 \text{ Å}^{-1}$ of all three proteins. The quality of the fit to the pattern obtained from the MD simulations is quantified as

$$\zeta = \sum_{n=0}^{N} \left( \ln \left( \frac{I_{MD}(q_n)}{I(q_n)} \right) \right)^2$$

and is presented in Table 3-1. Figure 3-16 displays the information in Table 3-1 as a bar graph. A comparison is also provided to calculations with CRYSOL and the optimization of the hydration shell density and the excluded volume to fit the experimental data. Using adjustable parameters can compensate for inherent deficiencies in the use of a uniform hydration layer but may also obscure errors in the molecular model used for the calculation. For our test set of 3 proteins, the radii of gyration ($R_g$) for the HyPred and CRYSOL calculations are within 0.5 Å of the MD determined values (Table 3-2). These $R_g$ are obtained from the second moment of $P(r)$, the electron pair distance distribution function as determined using the GNOM indirect Fourier transform procedure(71) and retaining scattering data over the range 0.06-0.6 Å $^{-1}$. 

69
Figure 3-14: Comparison between various hydration models.
Table 3-1: The difference between the specified scattering intensity and that calculated from the MD simulation, quantified as $\zeta$. The difference between the specified scattering intensity and the experimental, quantified as $\zeta$, is in parenthesis. Continued on next page.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Experiment</th>
<th>HyPred</th>
<th>CRY SOL</th>
<th>CRY SOL Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ub</td>
<td>0.0074</td>
<td>0.0011</td>
<td>0.070</td>
<td>0.0015 (0.0088)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.077</td>
<td>(0.0014)</td>
</tr>
<tr>
<td>HEWL</td>
<td>0.0066</td>
<td>0.0013</td>
<td>0.023</td>
<td>0.00091 (0.0082)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.040</td>
<td>(0.00091)</td>
</tr>
<tr>
<td>Mb</td>
<td>0.020</td>
<td>0.0068</td>
<td>0.075</td>
<td>0.0012 (0.016)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.079</td>
<td>(0.00094)</td>
</tr>
<tr>
<td>1CRC</td>
<td>0.0049</td>
<td>0.0094</td>
<td>0.034</td>
<td>0.012 (0.0054)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.032</td>
<td>(0.0050)</td>
</tr>
<tr>
<td>1DF4</td>
<td></td>
<td>0.00068</td>
<td>0.076</td>
<td>0.00020</td>
</tr>
<tr>
<td>1DXG</td>
<td></td>
<td>0.0012</td>
<td>0.22</td>
<td>0.00013</td>
</tr>
<tr>
<td>1F94</td>
<td></td>
<td>0.0021</td>
<td>0.019</td>
<td>0.00025</td>
</tr>
<tr>
<td>1HYP</td>
<td></td>
<td>0.0053</td>
<td>0.089</td>
<td>0.0010</td>
</tr>
<tr>
<td>1L2P</td>
<td></td>
<td>0.0060</td>
<td>0.022</td>
<td>0.00011</td>
</tr>
<tr>
<td>1TIF</td>
<td></td>
<td>0.00052</td>
<td>0.043</td>
<td>0.00054</td>
</tr>
<tr>
<td>1UOY</td>
<td></td>
<td>0.0023</td>
<td>0.029</td>
<td>0.00047</td>
</tr>
<tr>
<td>1US0</td>
<td></td>
<td>0.0010</td>
<td>0.033</td>
<td>0.0019</td>
</tr>
<tr>
<td>1VCC</td>
<td></td>
<td>0.0023</td>
<td>0.027</td>
<td>0.00054</td>
</tr>
<tr>
<td>1YZM</td>
<td></td>
<td>0.0013</td>
<td>0.058</td>
<td>0.00018</td>
</tr>
<tr>
<td>2DOB</td>
<td></td>
<td>0.0026</td>
<td>0.065</td>
<td>0.00072</td>
</tr>
<tr>
<td>2PTN</td>
<td></td>
<td>0.0077</td>
<td>0.042</td>
<td>0.0016</td>
</tr>
<tr>
<td>Protein</td>
<td>Experiment (Å)</td>
<td>MD (Å)</td>
<td>HyPred (Å)</td>
<td>CRY SOL (Å)</td>
</tr>
<tr>
<td>---------</td>
<td>----------------</td>
<td>--------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>Ub</td>
<td>12.40 ± 0.02</td>
<td>12.62 ± 0.07</td>
<td>13.15</td>
<td>13.21</td>
</tr>
<tr>
<td>HEWL</td>
<td>14.61 ± 0.04</td>
<td>14.44 ± 0.02</td>
<td>14.45</td>
<td>14.90</td>
</tr>
<tr>
<td>Mb</td>
<td>16.28 ± 0.04</td>
<td>16.61 ± 0.19</td>
<td>16.39</td>
<td>16.58</td>
</tr>
<tr>
<td>1CRC</td>
<td>12.89 ± 0.01</td>
<td>13.09 ± 0.08</td>
<td>13.11</td>
<td>10.71</td>
</tr>
<tr>
<td>1DF4</td>
<td>13.45 ± 0.28</td>
<td>13.46</td>
<td>14.73</td>
<td>62</td>
</tr>
<tr>
<td>1DXG</td>
<td>10.64 ± 0.15</td>
<td>10.31</td>
<td>11.66</td>
<td>36</td>
</tr>
<tr>
<td>1F94</td>
<td>11.25 ± 0.21</td>
<td>11.66</td>
<td>12.24</td>
<td>63</td>
</tr>
<tr>
<td>1HYP</td>
<td>12.41 ± 0.01</td>
<td>12.11</td>
<td>12.86</td>
<td>75</td>
</tr>
<tr>
<td>1L2P</td>
<td>9.93 ± 0.73</td>
<td>8.24</td>
<td>10.22</td>
<td>63</td>
</tr>
<tr>
<td>1TIF</td>
<td>14.21 ± 0.13</td>
<td>14.82</td>
<td>14.11</td>
<td>76</td>
</tr>
<tr>
<td>1UOY</td>
<td>10.39 ± 0.04</td>
<td>10.15</td>
<td>11.04</td>
<td>64</td>
</tr>
<tr>
<td>1US0</td>
<td>19.61 ± 0.02</td>
<td>19.64</td>
<td>19.68</td>
<td>314</td>
</tr>
<tr>
<td>1VCC</td>
<td>13.05 ± 0.07</td>
<td>12.93</td>
<td>13.76</td>
<td>76</td>
</tr>
<tr>
<td>1YZM</td>
<td>12.46 ± 0.11</td>
<td>12.35</td>
<td>12.59</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>2DOB</td>
<td>13.05 ± 0.03</td>
<td>13.04</td>
<td>13.56</td>
<td>82</td>
</tr>
<tr>
<td>2PTN</td>
<td>16.89 ± 0.05</td>
<td>16.65</td>
<td>17.41</td>
<td>230</td>
</tr>
<tr>
<td>2ZQE</td>
<td>12.33 ± 0.04</td>
<td>12.13</td>
<td>12.94</td>
<td>80</td>
</tr>
<tr>
<td>3G19</td>
<td>13.34 ± 0.13</td>
<td>13.61</td>
<td>14.24</td>
<td>84</td>
</tr>
<tr>
<td>3HGL</td>
<td>12.28 ± 0.03</td>
<td>12.32</td>
<td>12.97</td>
<td>77</td>
</tr>
<tr>
<td>3LE4</td>
<td>12.54 ± 0.11</td>
<td>12.52</td>
<td>13.28</td>
<td>55</td>
</tr>
</tbody>
</table>

Table 3-2 (continued)

*These $R_g$s are obtained from the second moment of $P(r)$, the electron pair distance distribution function as determined using the GNOM indirect Fourier transform procedure, and retaining scattering data over the range 0.06-0.6 Å⁻¹. The error bars for the experimental $R_g$s are obtained from GNOM. The error bars for the MD simulations are obtained by splitting the MD simulations into three pieces and calculating the $R_g$s of the individual pieces.*
Figure 3-15: Agreement between the MD simulation with either the experimental data or the various hydration models. The $y$ axis is the fit of a scattering pattern to the scattering pattern calculated from the MD simulation, as defined in equation 3-8.

The $P(r)$ functions are calculated from the intensity distributions determined with the HyPred model. Generally, the computed $P(r)$ are very similar for the different treatments of the solvent contributions (Fig. 3-16). The $P(r)$ curves for the data in the range of 0.06-2.4 Å$^{-1}$ exhibit relatively high frequency oscillations that correspond to features obtained when $P(r)$ is calculated directly from the atomic coordinates for the structure, providing validation for the approach and the indirect Fourier transform used to calculate $P(r)$. Figure 3-17 displays the $P(r)$
calculated directly from the structure of HEWL with the hydration layer calculated using HyPred and the \( P(r) \) calculated from the direct Fourier transformation of the scattering intensity calculated from the MD simulation of HEWL for \( q \) out to extremely high \( q = 30 \text{ Å}^{-1} \), which is essentially infinity. The \( P(r) \) curves calculated in two different ways feature the same major oscillations.

Figure 3-16: \( P(r) \) curves calculated using various data ranges. The \( P(r) \) for Ub, HEWL, and Mb calculated using GNOM(71) with data from \( q=0.06 \text{ Å}^{-1} \) out to 0.3 (top), 0.6 (middle) and 2.4 (bottom) Å\(^{-1}\). The experimental data (black), MD (red), HyPred (blue) and CRYSOl (green) are shown.
Figure 3-17: Comparison between a direct Fourier transformation of the scattering calculated from the MD simulation of HEWL out to $q=30$ Å$^{-1}$ and the $P(r)$ curve calculated from the HEWL crystal structure hydrated by HyPred.

3.5 Conclusion

The ability of the parameter free HyPred continuum-hydration model to produce scattering patterns that closely match the explicit atom solvent calculations demonstrates that SWAXS patterns can be calculated without MD simulations using an inhomogeneous continuum model of the hydration shell. This agreement represents a strong validation of the HyPred method for the calculation of the solvation layer of proteins. This also implies that solvation of protein surfaces is a highly local phenomenon that can be accurately predicted from the relative location of atoms in the immediate vicinity. The success of HyPred also implies that the
density correlations and fluctuations within the hydration shell do not significantly contribute to the SWAXS pattern for \( q < 1.3 \ \text{Å}^{-1} \), while inhomogeneities in the hydration shell do make an impact. This claim follows because the SWAXS patterns calculated from the MD simulations, which include contributions from density correlations, closely match the SWAXS patterns calculated from HyPred. We assess the impact of the improved description of the hydration shell and the improved method of correcting for bulk solvent contributions and find that both are significant. Our methods produce significantly more accurate SWAXS patterns than those obtained using the program CRYSOL with default parameters. While the computations using HyPred may currently be too slow for use “on-the-fly” during a folding or docking algorithm, they can be employed to filter computationally generated structures. Thus, our approaches should aid in the modeling of multi-domain protein structures and complexes.

The HyPred methodology can be extended to enable calculation of the hydration of nucleic acids and membrane surfaces and, hence the SWAXS patterns for nucleic acid-protein complexes and membrane proteins in lipid vesicles or bilayers. HyPred may also be employed to estimate water densities surrounding proteins and thereby aid in crystallographic structure refinement. The HyPred based software presented in this paper is accessible through a web server at http://godzilla.uchicago.edu/cgi-bin/jouko/waxs.cgi. A web server for performing SWAXS calculations from MD simulations is accessible at http://godzilla.uchicago.edu/cgi-bin/jouko/md_waxs.cgi.
4. Modeling the hydration layer around nucleic acids

4.1 Abstract

Hydration plays an integral role in many biological processes. The HyPred theory for predicting the solvent density near the surface of a protein to atomic level precision is extended to predict the hydration layer surrounding highly charged RNA molecules. Hydration densities are extracted for seven RNAs from all-atom explicit solvent molecular dynamics simulation. Large differences emerge in the hydration patterns between the A and B-forms and near CG versus UA/TA base pairs, while the hydration layers surrounding RNA and DNA structures are fairly similar. The RNA simulations are used to calculate a set of universal, atom dependent proximal radial distribution functions. The HyPred model and the proximal distribution functions for a given RNA molecule are tested in their ability to reproduce the hydration layer from the simulations for that RNA as well as those for other RNA molecules. HyPred quickly and accurately predicts the hydrations shells around nucleic acids without the need for new costly MD simulations.

4.2 Introduction

Solvation is essential for all life on earth. The solvent strongly influences a wide range of biological processes, including protein and nucleic acid folding (75), stability (20, 21, 76) and dynamics (22, 77), enzymology, including transition state stabilization (23, 78), binding (24, 25,
diffusion, and electrostatic interactions (26, 81), to name a few. In addition, the presence of solvent surrounding biological macromolecules affects many biophysical measurements, including NMR spectroscopy (28-30, 82), X-ray crystallography (31, 32, 83), small/wide angle X-ray scattering (SWAXS) (33, 84)(Virtanen 2010), and neutron diffraction (34, 35, 85). A rapid and accurate theory to predict the solvent density around biological macromolecules would facilitate the unbiased interpretation of data from these experimental techniques, as accomplished in a previous paper (17) for SWAXS scattering from proteins in solution. The same general model used for proteins is extended to describe the hydration layer of the more highly charged RNA molecules.

Water is essential for the biological activity of RNA molecules. Water molecules participate directly in ribozyme catalysis (78) as well as by coordinating ions and RNAs (86). An early MD simulation of RNA with and without water demonstrates that water is essential for RNA stability (76). More recently, Pande has performed many MD simulations for a small RNA molecule using different water models to determine the role that water plays in RNA folding (75). Even long range RNA solvent interactions are important in imparting stability to folded RNA (87, 88). Similarly, the ordering of water dipoles around proteins and DNAs contribute significantly to the protein-DNA binding process (89).

Many previous studies address the hydration of biological macromolecules. Pettitt has studied the hydration of both proteins (37) and nucleic acids (7) using a similar but less detailed approach as the HyPred theory developed by us (17). Other theoretical approaches to DNA
hydration include the use of integral equations methods by Pettitt (90) and Pande (11). Hummer has used multibody correlation functions to predict solvent densities around DNA. Dill has developed a semi-explicit model of water in which distribution functions for the average water density and dipole moments are first calculated for solvated single atoms with various partial charges and van der Waals radii. Then, dipoles representing water molecules are placed in the hydration layer according to the calculated potentials (91). While crystallography provides a probe of RNA hydration, even high resolution crystal structures cannot generate an accurate or complete picture of the hydration, though there have been attempts at predicting crystallographic waters using crystallographic structures (12-15).

Our theory HyPred (17), which has previously been used to predict the hydration layer around proteins, is extended to describe the hydration layer surrounding nucleic acids in a manner that removes the difficulties with previous attempts as follows. Prior theoretical studies of RNA hydration allow the RNA molecule to be mobile (92), a feature shown by us to introduce artifacts due to the smearing of the hydration shell because the solvent density at a particular site near the surface is mostly determined by the percentage of the time that the site is occupied by the RNA molecule (17) rather than the inherent molecular structure of the hydration layer. Some studies just consider the local hydration around individual base pairs to mitigate this problem (93). We completely eliminate the problem by using MD simulations in which the RNA molecules are held immobile.
Hydration densities are extracted from the MD simulations by superimposing a cubic grid on the MD simulation box and calculating the average solvent electron density in each cube. We analyze the hydration around different base pairs and different RNA conformations. A set of proximal radial distribution functions (pRDFs) is calculated from the MD simulations with the goal of predicting the solvent electron density around nucleic acids for which MD simulations have not been performed. The pRDFs are largely able to recapitulate the Å level pattern of density fluctuations around the RNAs. An analysis of the hydration pattern around A- and B-form RNA and DNA exhibits the presence of differences particularly in the major and minor grooves.

4.3 Results

4.3.1 Solvent density maps

The solvent density maps for seven different immobile RNA molecules are determined from all-atom, explicit solvent MD simulations using the program NAMD (94). Additional simulations are performed for A-form and B-form RNA with the same sequence as 1A1T and for DNA analogs of the A-form and B-form RNA structures. Despite the highly disfavored nature of B-form RNA and TA pairs in the A-form of DNA, MD simulation of B-form RNA and A-form DNA containing TA pairs are used for comparison with the other structures, so that the influence of structure and chemical composition can be more easily analyzed. Otherwise, it is difficult to assess whether a large difference between the hydration of A-form UA base pairs in RNA and that for B-form TA base pairs of DNA emerges from differences in A-form and B-form
structures, differences in uracil and thymine nucleotide bases, or the presence of a hydroxyl group on RNA.

Four different ionic conditions ranging from 20 mM to 1 M NaCl are tested, and the results are very similar for the four cases (Figures 4-1 and 4-2). As found in our previous studies (17) comparisons of simulations for both mobile and immobile proteins, the accurate calculation of the solvent electron density surrounding a solute necessitates that only the water molecules be permitted to move during the MD simulations. The average solvent electron densities, obtained from simulations of durations 32-64 ns, are discretized into a lattice of (0.5 Å)³ cubes for further analysis.

![Figure 4-1: Cross sections showing the hydration shell density of 1A1T at 20 mM NaCl, 145 mM NaCl, 145 mM NaCl and 10 mM MgCl₂, and 1 M NaCl.](image)

Figure 4-1: Cross sections showing the hydration shell density of 1A1T at 20 mM NaCl, 145 mM NaCl, 145 mM NaCl and 10 mM MgCl₂, and 1 M NaCl.
Figure 4-2: pRDFs calculated from MD simulations performed at 20 mM and 1 M NaCl.

Figure 4-3 displays a cross section of the solvent density surrounding a CG pair in an A-form RNA with the same sequence as 1A1T. The cross section exhibits regions of high density near the sugar edge (G)N2-H2, (G)O2'-H2', (G)O1P, (G)O2P, (G)N7, (G)O6, (C)N4-H42, (C)O1P, and (C)O2P atoms. All the regions of high density in the cross section appear near either hydrogen bond donors or acceptors. Many regions of high and low density are also evident in the major groove, far away from any of the atoms displayed. Such regions in the major groove of the CG pair in A-form RNA emerge because of the presence of the upstream and downstream oxygen atoms of the phosphate backbone located one half turn away (Figure 4-4).

4.3.2 Comparison between A and B-forms

Despite the highly disfavored nature of B-form RNA, we perform an MD simulation of the (immobile) B-form RNA for comparison with the A-form. The hydration pattern around the B-form of RNA displayed in Figure 4-3b differs significantly from that for A-RNA. When viewed in cross section, the inner surfaces of the hydrations shells of the CG base pair in A-form and B-
form RNA are oval and crescent shaped, respectively. The B-form crescent contains a deep minor groove and a shallow major groove. The cross section of the CG pair in B-form RNA is missing the regions of high and low density corresponding to those observed in the major groove of the CG pair in A-form RNA that are associated with upstream and downstream atoms of the

Figure 4-3: Comparison of the hydration of CG pairs between A-form and B-form nucleic acids and between DNA and RNA. A) Solvent density around a CG pair in A-form RNA. B) Solvent density around a CG pair in B-form RNA. C) Solvent density around a CG pair in A-form DNA. D) Solvent density around a CG pair in B-form DNA. In all figures the major grooves are at the top.
Figure 4-4: Cross section of the same CG base pair as displayed in Figure 4-3, with the full RNA structure. Regions of high density located in between the oxygen atoms of the phosphate backbone, while regions of low density appear in front of the oxygen atoms of the phosphate backbone because the phosphate groups exclude water molecules from those regions by van der Waals forces.

phosphate backbone (Figure 4-3b). This difference arises because the phosphate groups of the opposing strands that flank the major groove are further apart in the B-form.

The region of high density near the sugar edge (G)N2-H2 atom in A-form RNA is absent from the hydration shell of the B-form of RNA. Instead, a region of high solvent density is common to the (G)N3 and (G)O2’-H atoms in the B-form RNA structure, whereas no
corresponding region of high solvent density appears near the (G)N3 atom in the A-form of RNA. The region of high density appearing near the sugar edge (G)N2-H2 atom in the A-form of RNA is less dense than in the A-form of DNA, possibly because the (G)O2’-H2’ atom contributes indirectly to the region of high density near the sugar edge (G)N2-H2 atom through the region of high density near the (G)O2’-H2’ (i.e., the sugar edge (G)N2-H2 atom in RNA binds a water molecule that is also bound to another water molecule that, in turn, is bound to the (G)O2’-H2’ atom). With the exception of the high and low density regions caused by the proximity of the oxygen atoms of the phosphate backbone, the solvation layers around the major grooves of A-RNA and B-RNA are more similar to each other than the solvation layers around the minor grooves.

4.3.3 Comparison between A-forms of RNA and DNA

The cross section of a CG pair in A-form DNA, displayed in Figure 4-3c, is very similar to the cross section of the CG pair of A-form RNA. The only significant difference is that the region of high density near the (G)O2’-H2’ atom in the cross section of the A-form of RNA is absent from the cross section of the A-form of DNA because DNA lacks the (G)O2’-H2’ atom. The cross section of a CG pair of B-form DNA, displayed in Figure 4-3d, is even more similar to its RNA analog than the cross section the CG pair of A-DNA is similar to its RNA analog because the 2’ hydroxyl group on the ribose sugars are pointed inwards towards the nucleotide bases and, therefore, do not modify the hydration layer of B-form RNA structures.
A few regions of high density are present in the cross section of a UA pair from the A-form RNA structure displayed in Figure 4-5a. With the exception of one region of high density near the Hoogsteen edge (A)N6-H2 atom, regions of high density in the cross section of the UA pair in A-form RNA are generally associated with the backbone. Similar to the regions of high and low density in cross sections of the CG pairs of A-form RNA and A-form DNA associated with the upstream and downstream oxygen atoms of the phosphate backbone, regions of high and low density appear in the cross sections of the UA pair in A-form RNA and TA pair in A-form DNA. The hydration patterns around the UA pair in A-form RNA and the TA pair in A-form DNA are similar, except the region of high density near the (A)O2’-H2’ atom present in the cross section of the UA pair in A-form RNA is absent from the cross section of the TA pair in A-form RNA because DNA lacks the (A)O2’-H2’ atom. In addition, the region of high density near the (U)O2P atom in the cross section of the UA pair in A-form RNA is missing in the cross section of TA pair in A-form DNA, possibly due to the influence of the extra methyl group on thymine that is absent from uracil. The region of high density near Hoogsteen edge (A)N6-H2 atom present in the UA pair from A-form RNA is absent from the cross section of the TA pair of A-form DNA.

Despite the highly disfavored nature of TA pairs in the A-form of DNA, we perform an MD simulation of A-form DNA for comparison with the A-form RNA. Another region of high density in the cross section of the UA pair in the B-form RNA structure appears near the (A)N3 atom. The cross sections of the UA pair in B-form RNA and TA pair in B-form DNA are very similar. Since the density scale is asymmetrical and a small decrease relative to bulk density
leads to a cube being represented in blue but a cube with a small increase relative to bulk density need not be displayed in red, many cubes in the bulk solvent appear blue.

More cross sections through CG pairs are displayed in Figures 4-6a-e, and Figures 4-6f-h exhibit additional examples of cross sections through UA pairs. Figures 4-6a-c display cross sections from A-form base pairs, Figures 4-6d,f present cross sections from TA-form base pairs, and Figures 4-6e,g,h provide cross sections from irregular structures.

Figure 4-5: Comparison of the hydration of UA pairs between A-form and B-form RNA and between TA pairs of A-form DNA and B-form DNA. A) Solvent density around a UA pair in A-form RNA. B) Solvent density around a UA pair in B-form RNA. C) Solvent density around a TA pair in A-form DNA. D) Solvent density around a TA pair in B-form DNA.
A) GC pair in A-form RNA
B) GC pair in A-form RNA
C) GC pair in A-form RNA
D) GC pair in TA-form RNA
E) GC pair in irregular RNA
F) AU pair in TA-form RNA
G) AU pair in irregular RNA
H) AU pair in irregular RNA

Figure 4-6: More examples of cross sections of hydration shells.
4.3.4 High density regions

We now focus on regions of high solvent electron density, defined as any cube with a density greater than $10 \text{ e}^{-3} \text{Å}^{-3}$, a level 30-fold higher than bulk solvent ($0.334 \text{ e}^{-3} \text{Å}^{-3}$). The occurrence of such a high density is only possible because the cubes are much smaller than water molecules, which occupy approximately 90 cubes on average. A density of $10 \text{ e}^{-3} \text{Å}^{-3}$ in a $(0.5 \text{ Å})^3$ cube would occur if the cube were occupied by an oxygen atom of water 16% of the time.

Many high density points (red) around 1A1T are located in the major and minor grooves (Figures 4-7a,b). In addition, many such regions are present near the phosphate groups of the backbone and the 2’ hydroxyl groups of the ribose sugars. Some regions of high density are located near bases. For example, regions of high solvent density appear near the (G)N2-H2 atoms that are available for hydrogen bonding. Unpaired bases that extend into the solution tend to have regions of high density near their edges, as illustrated in Figures 4-7a,b. Cross sections of the hydration shell densities are displayed in Figures 4-7c,d. Many regions of high density are naturally followed by regions of low density because of hard core excluded volume interactions.
Figure 4-7: Comparison between the hydration shells obtained from MD simulations and predicted by HyPred. Regions of high hydration density, from MD simulations, are displayed in red in A and B. Cross sections of the hydration shells, from MD simulations, are displayed in C and D. Cross sections of the HyPred predictions are shown in E and F.

4.3.5 Proximal radial distribution functions

A general method for calculating the hydration layer around arbitrary RNA molecules begins with pRDFs calculated from MD simulations for each type of atom in RNA molecules. HyPred then uses these pRDFs to calculate the hydration for any arbitrary RNA molecule without further simulations. To calculate the pRDFs, we determine the scaled van der Waals
surface by surrounding all solute atoms with spheres of radii $r=0.53r_{vdw}$, where $r_{vdw}$ is the standard van der Waals radius, and then every cube is assigned to the solute atom with the closest surface. The distances between each cube and the atom with nearest surface is then employed to calculate the pRDFs for the solute’s surface atoms as described in Methods.

The pRDFs are calculated by grouping cubes according to the atom type of the nearest solute atom and the distance to the nearest solute atom and by finding the average density as a function of the distance from the nearest solute atom and of the atom type of the nearest solute atom. Two different sets of pRDFs are calculated. One set groups heavy atoms according to element, while hydrogen atoms are further subdivided according to the atom to which the hydrogen atoms are bonded. In the other more detailed set, every unique atom forms a separate type. Further refinement involves inclusion of the dependence of the pRDFs upon the second nearest neighbor, the complement of the angle formed by the line connecting the cube and the nearest solute atom and the line connecting the nearest solute atom and the atom to which it is bonded (Figure 3-5), and the concavity of the region in which the cube is located. First, the pRDFs calculated from the MD simulation for an RNA molecule are used to reconstruct that RNA’s hydration shell, and then the average pRDFs calculated from other RNA molecules are used to predict the same hydration shell.

Four sample pRDFs are presented in Figure 4-8a. Of the four pRDFs, the one for the (G)N2-H2x atom type contains a peak with the greatest electron density, followed by the pRDFs for (C)PO and (U)O4 atom types. Hydrophilic atoms yield pRDFs with higher peaks than hydrophobic atoms. The pRDF for the hydrophobic (G)C6 atom contains the lowest and most
distant first maximum. The pRDFs displayed in Figure 4-8a illustrate the general trends that the regions of highest density in the solvent preferentially reside near the phosphate backbone and the exposed hydrogen bond donors and acceptors of the bases. The pRDFs for the hydrogen bond acceptors (C)PO and (U)O4 contain small peaks at 1.7 Å and 1.9 Å, respectively. The first peak is due to the hydrogen atoms, while the second larger peak is mainly due to the oxygen atoms. The pRDF of the hydrogen bond donor (G)N2-H2x contains the closest first maximum at 2.2 Å.

**Figure 4-8: pRDFs of RNA and protein atom types.** pRDFs of various RNA atom types are displayed in A. Comparisons between pRDFs of RNA and protein atom types are displayed in B-E. In F all atom types are combined into one and pRDFs are calculated for RNA molecules and proteins. In B-F the pRDFs of the RNA atom types are displayed in black and the pRDFs of the RNA atom types are displayed in grey.
Previous studies of the hydration layer surrounding proteins demonstrate that an improved description of the hydration layer emerges when the cubes are partitioned into groups with higher specificity (17). Hence, we also label each cube by the solute atom, termed the “second nearest”, whose scaled van der Waals surface is the next nearest to each cube. The cubes are also labeled by the (discretized) complement of the angle, termed the θ angle, formed by the line between the center of the cube and its assigned atom and the line between the assigned atom and the atom to which it is bonded (Figure 3-5). If more than one atom is bonded to the solute atom nearest to the cube, we use the first bonded atom listed in a psf file. The angular dependence is more important when only one atom is covalently bonded to the nearest solute atom, e.g. hydrogen atoms and carbonyl oxygen atoms. Finally, we distinguish the cubes according to whether they lie in convex or concave regions of the RNA.

Figure 4-9 displays two pRDFs with dependencies on the second nearest neighbor, the θ angle, and the surface concavity. The pRDF displayed in Panel a (and designated as \( \text{pRDF}_{\text{NH,concave,CH,0-15}} \)) is calculated from the subset of cubes that are nearest to a hydrogen atom bonded to a nitrogen atom, are located in a concave region, have a hydrogen bonded to a carbon atom as the second nearest neighbor, and have the θ angle less than 15°. The peak in the \( \text{pRDF}_{\text{NH,concave,CH,0-15}} \) is very high, with \( \rho=13.6 \text{ e}^{-\text{Å}^{-3}} \) or ~40 times the bulk density (0.334 e Å⁻³). The pRDF also contains a strong minimum of only 0.08 e Å⁻³ because the presence of the water molecules responsible for the peak in the \( \text{pRDF}_{\text{NH,concave,CH,0-15}} \) excludes water molecules from nearby regions. The pRDF displayed in Figure 4-9b (and designated as \( \text{pRDF}_{\text{CH,concave,NH,15-30}} \)) is calculated from the subset of cubes that are nearest to a hydrogen atom...
bonded to a carbon atom, are located in a concave region, have a hydrogen atom bonded to a carbon atom as the second nearest atom, and have the θ angle between 15° and 30°.

$p_{\text{RDF}}_{\text{CH,concave,NH,15-30}}$ contains a significantly higher peak (1.75 e Å⁻³) than the average pRDF for hydrogen atoms bonded to carbon (0.72 e Å⁻³) because of the proximity of the hydrogen atom bonded to nitrogen.

Figure 4-9: pRDFs categorized according to the concavity of the region in which the cube is located, the second nearest neighbor, and the angular dependence. A) The pRDF calculated from the subset of cubes which are nearest to a hydrogen atom bonded to a carbon atom, are located in a concave region, have a hydrogen bonded to a carbon atom as the second nearest neighbor, and where the angle formed by the line connecting the cube and the nearest atom and the line connecting the hydrogen and nitrogen atom is less than 15°. B) The pRDF calculated from the subset of cubes which are nearest to a hydrogen bonded to a carbon atom, are located in a concave region, have a hydrogen atom bonded to a carbon atom as the second nearest atom, and where the angle formed by the line connecting the cube and the nearest atom and the line connecting the nearest atom and the nitrogen atom to which it is bonded is between 15° and 30°.
4.3.6 Comparison to protein pRDFs

A simple visual comparison between the protein and RNA pRDFs in Figure 4-8b-e appears to suggest that the hydration shells of proteins and RNAs are similar. However when all atoms are combined into one atom type so that one pRDF is calculated each for proteins and for RNAs, the averaged pRDFs are very different (Figure 4-8f). The hydration density around RNAs is significantly higher ($0.85 \text{ e}^{\cdot} \text{Å}^{-3}$) than for the proteins ($0.58 \text{ e}^{\cdot} \text{Å}^{-3}$) because of the greater charge density and polar atom types at the surface.

The pRDFs of some of the RNA atom types are compared with the previously determined pRDFs for similar protein atom types (17). For example, the Asp Oδ atom of proteins is taken to be the analog of the (C)PO atom of RNA. The partial charges of (C)PO and Asp Oδ atoms are very similar in the force fields used, -0.776 and -0.76, respectively. Likewise, the van der Waals radii are similar, 1.7 and 1.66 Å, respectively. The two pRDFs are quite similar (Figure 4-8b), except the pRDF for Asp Oδ contains a higher peak than the pRDF for (C)PO. The difference is partly due to a difference in the oxygen atoms’ geometry. The Cyt’s PO-P-PO angle is $140^{\circ}$ and the P-PO bond length is 1.48 Å, while the Asp’s Oδ-Cδ-Oδ angle is $124^{\circ}$ and the Cδ-Oδ bond length is 1.26 Å. The colocalization of the Asp’s two oxygen atoms makes them more hydrophilic, although other factors such as the size of water molecules may also play a role. The preceding example demonstrates that the charge and van der Waals radius of an atom are insufficient to accurately determine the pRDF. The surrounding environment also influences pRDFs.
The influence of the environment is illustrated in Figure 4-8c, which compares the pRDF of the (U)O4 atom with the pRDF of the carbonyl oxygen atom of methionine. The peak for the methionine is higher despite the opposite ordering of the absolute values of the partial charges, -0.51 versus -0.58. The van der Waals radii of the carbonyl oxygen of methionine and (U)O4 are almost identical (1.72 and 1.66 Å, respectively). Since the partial charge of the (U)O4 atom exceeds that of the carbonyl oxygen of methionine, the lower density in the pRDF for (U)O4 must arise from other factors in the surrounding environment.

The previous example is not unique. Figure 4-8d illustrates that the pRDFs of the (G)N2-H2x and Gln Hε2x atoms are very similar despite differences in their partial charges, 0.44 vs. 0.32, respectively, and van der Waals radii, 0.6 and 0.22 Å, respectively. The pRDFs for the (G)C6 atom and the carbonyl carbon of alanine in Figure 4-8e resemble each other, except the peak in the pRDF of the (G)C6 atom exceeds the peak of the pRDF of the carbonyl carbon of alanine due to more hydrophilic environment around (G)C6. The partial charges and van der Waals radii of the (G)C6 atom and alanine carbonyl carbon atom are close, 0.477 and 0.5 and 1.908 and 2.0 Å, respectively. Both of the (G)C6 and carbonyl carbon of alanine pRDFs contain more statistical noise than the other pRDFs because the (G)C6 atom and carbonyl carbon atoms tend to be buried and, consequently, are rarely exposed to solvent, leading to poor statistics.

4.3.7 Reconstructions

Hydration shell densities of the individual RNA molecules are reconstructed by using the pRDFs from the MD simulations and reversing the procedure followed for calculating the
pRDFs. Examples of the reconstructions are displayed in Figures 4-7e,f to facilitate comparisons with the original densities obtained from the MD simulations. All figures displaying HyPred reconstructions for particular RNA molecules are calculated using the average of the pRDFs evaluated from the simulations for the other RNA molecules. The overall pattern of density fluctuations is reproduced well. A detailed examination, however, indicates that the regions of high density in the reconstructions are not always as dense as the regions of high density in the MD simulations, and the regions of low density in the reconstructions sometimes are denser than the regions of low density in the MD simulations.

Figs. 4-7c,e contain regions of high density near the (G207)O1P, (G207)O2P, (G217)N2-H22, and (C219)O2P atoms in both the cross section of the hydration shell obtained from the MD simulation and the HyPred prediction. These regions of high density are much less pronounced and more spread out in the HyPred reconstructions. A region of high density near the (A211)N1 atom in the cross section of the hydration shell obtained from the MD simulation in absent from the reconstruction. Figures 4-7d,f exhibit regions of high density near the (C2)N2-H42, (G12)N7, (A11)O2P, (A8)N6-H61, (G6)O1P, (U5)O2P, (C4)O3’, and (C2)O2’ atoms in both the cross section of the hydration shell obtained from the MD simulation and the HyPred prediction. Many of the regions of high density in the HyPred prediction are faint and smeared out over a large volume, with the region of high density near the (C2)N2 atom being the only region of high density that is clearly visible, in the HyPred reconstruction. A region of low density appears at the bottom left of the cross section of the MD simulation of 1HJ1 in Figure 4-7d, but is absent in the HyPred prediction. In addition to reproducing the initial depletion layer
and the layer of greater than bulk density, HyPred reproduces the second depletion layer which follows the layer of greater than bulk density. The reconstructions produced using the enhanced version of HyPred with conditionals upon second neighbor, θ angle, and surface geometry, are not as smooth as the MD simulations because of the sparseness of the data when the MD data are partitioned into more pRDFs. Simulations for additional RNA molecules would help smooth the predictions, but the available pRDFs are sufficient for present purposes.

When using pRDFs determined from the MD simulation for an RNA molecule in the reconstruction for the same RNA, the accuracy of the HyPred procedure as measured by the real space R factors, defined in the Methods section, ranges from 0.085 to 0.117. When HyPred is employed to predict the hydration layers using the average of the pRDFs from the simulations for the other RNAs as test of the universality and transferability of the pRDFs, the R factors range from 0.114 to 0.145, about 0.03 higher than when using the pRDFs for the same RNA molecule.

We test two different sets of pRDFs. One set of pRDFs, termed the coarse grained set, categorizes heavy atoms by their element type (e.g., C, N, O, S...), and subdivides hydrogen atoms according to the element to which they are bonded (e.g., CH, NH, OH). In the other set of pRDFs, termed the fine grained set, every unique atom in every nucleotide base and in the backbone is its own atom type (Table 2-1). The R factors, obtained using the coarse grained set of atom types and the RNA molecules’ own pRDFs, are very similar to the R factors obtained using the coarse grained set of atom types and the average of the pRDFs obtained from the MD simulations of the other RNA molecules. The R factors determined using the coarse grained set
of pRDFs are about the same as the R factors obtained for the fine grained set of pRDFs. Including simulation data for more RNAs to calculate the pRDFs should improve the R factors generated with the fine grained set of pRDFs, but the R factors obtained using the coarse grained set of pRDFs would not drop significantly because the R factors emerging from the coarse grained set are about the same whether the RNA molecules own pRDFs are used or the average of the other RNAs’ pRDFs are used and because the R factor from the RNAs own pRDFs is a lower bound for the R factor with the averaged pRDFs.

The quality of all reconstructions, summarized in Table 4-1, are comparable to those found for proteins. However, the shortest RNA simulation at 32 ns in duration (Table 4-1), and exceeds the 9 ns used for the proteins. Longer simulations reduce the statistical noise and diminish the R factors. Using only the first 9 ns seconds of the RNA simulations produces R factors for the RNA molecules exceeding those for the proteins. Comparisons of the predictions for RNA and proteins are displayed in Figure 4-10.
A) A cross section of the hydration shell density of 1A1T calculated from an MD simulation.

B) A cross section of the hydration shell density of 1UBQ calculated from an MD simulation.

C) A cross section of the hydration shell density of 1A1T predicted by HyPred. R=0.13.

D) A cross section of the hydration shell density of 1UBQ predicted by HyPred. R=0.12

Figure 4-10: A comparison of the hydration shells of an RNA molecule and protein as computed from MD simulations and by HyPred.
<table>
<thead>
<tr>
<th>RNA</th>
<th>Self</th>
<th>Universal</th>
<th>Simulation Length (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>RMSD</td>
<td>R^*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R</td>
</tr>
<tr>
<td>1A1T</td>
<td>0.102</td>
<td>0.817</td>
<td>0.316</td>
</tr>
<tr>
<td></td>
<td>(0.127)</td>
<td>(1.05)</td>
<td>(0.418)</td>
</tr>
<tr>
<td>1E7K</td>
<td>0.101</td>
<td>0.746</td>
<td>0.332</td>
</tr>
<tr>
<td></td>
<td>(0.122)</td>
<td>(0.896)</td>
<td>(0.423)</td>
</tr>
<tr>
<td>1G70</td>
<td>0.099</td>
<td>0.697</td>
<td>0.291</td>
</tr>
<tr>
<td></td>
<td>(0.125)</td>
<td>(0.963)</td>
<td>(0.386)</td>
</tr>
<tr>
<td>1HJI</td>
<td>0.085</td>
<td>0.566</td>
<td>0.286</td>
</tr>
<tr>
<td></td>
<td>(0.109)</td>
<td>(0.809)</td>
<td>(0.389)</td>
</tr>
<tr>
<td>1JBR</td>
<td>0.115</td>
<td>0.847</td>
<td>0.344</td>
</tr>
<tr>
<td></td>
<td>(0.132)</td>
<td>(0.967)</td>
<td>(0.417)</td>
</tr>
<tr>
<td>1S03</td>
<td>0.117</td>
<td>0.886</td>
<td>0.330</td>
</tr>
<tr>
<td></td>
<td>(0.138)</td>
<td>(1.05)</td>
<td>(0.411)</td>
</tr>
<tr>
<td>2A9X</td>
<td>0.104</td>
<td>0.770</td>
<td>0.320</td>
</tr>
<tr>
<td></td>
<td>(0.128)</td>
<td>(0.973)</td>
<td>(0.421)</td>
</tr>
</tbody>
</table>

Table 4-1: The R factors for the HyPred predictions with respect to the MD simulations for the RNA molecules. Values in parenthesis are obtained with the coarse grained representation. Self refers to using an RNA’s own pRDFs to reconstruct its hydration shell. Universal refers to the average pRDFs of the other RNA molecules.

4.4 Discussion

We extend the accurate HyPred model for rapidly calculating the solvent density around a protein to predict the solvation layer surrounding RNAs with comparable accuracy and without the need for additional costly MD simulations. Both protein and RNA molecules have
1-2 Å size regions of high and low solvent density near the exposed hydrogen bond donors and acceptors. The hydration layer around RNA exhibits higher density along the phosphate backbone. HyPred yields slightly improved predictions of the hydration layer when account is taken of the second nearest neighbor, orientation, and the concavity of the surface. Since HyPred only classifies regions as either convex or concave, improvements could accrue by replacing this binary classification with a quantification of the local curvature of the solute. Additional improvements could be made by studying regions near a solutes surface, where a water molecule could form multiple hydrogen bonds with the solute.

Future applications will include the study of protein-RNA complexes and charge radial distribution functions, which could enable the calculation of the electrostatic contribution and the van der Waals potentials to the hydration energy. Other possibilities include predicting the distribution of water orientation around proteins and RNAs, the residency times of water molecules, and the number of hydrogen bonds between water molecules and the effects of temperature. A website for performing HyPred calculations can be found at http://godzilla.uchicago.edu/cgi-bin/jouko/HyPred.cgi.

4.5 Methods

4.5.1 MD simulations

All-atom explicit solvent MD simulations, employing NAMD(94) and the AMBER99(95) force field, are performed for the RNA portions of 1A1T(96), 1E7K(97), 1EHZ(98), 1G70(99), 1HJI(100), 1JBR(101), 1S03(102), and 2A9X(103). Non-canonical nucleotide bases are mutated
to their canonical analogs (e.g., pseudouracil nucleotide bases are mutated to uracil nucleotide bases). In addition, X3DNA(104) is used to create A and B forms of 1A1T whose hydration layers are determined from MD simulations. The A and B-forms of 1A1T, created using X3DNA, are also converted to DNAs, and MD simulations are performed for these DNA molecules. Four simulations are performed made for each RNA molecule, each with a different concentration of ions, namely, 20 mM NaCl, 145 mM NaCl, 145 mM NaCl and 10 mM MgCl₂, and 1 M NaCl. The hydration layers computed separately for each of the four different ionic concentrations are found to be nearly independent of the ionic conditions (Figures 4-1 and 4-2). Additional calculations using an average of the simulations for all four ionic conditions provide better statistics and, thus, are used for all the results reported in the main body of this article.

Protonation states are assigned assuming standard states at pH = 7. All simulations have box sizes whose dimensions extend 15 Å beyond the RNA molecule on all sides. Table 4-1 summarizes the durations of the simulations of the RNAs. The MD simulations of the canonical A-form and B-form RNA simulations are 4 ns in duration, and the MD simulations of the DNA molecules are 8 ns in duration. The density of the bulk solvent is determined in each case from all water molecules situated at distances exceeding 10 Å from the solute. If the computed electron density of the bulk water is either less than 0.333 e⁻ Å⁻³ or greater than 0.335 e⁻ Å⁻³, the simulation box size is adjusted to produce the correct bulk density.

Energy minimization and equilibration proceed in several stages. The solvent is first energy minimized for 200,000 steps; then, with all atoms in the nucleic acids immobilized, the temperature is initially set at 10 K and then raised by 10 K every 100 ps until reaching the final
temperature of 4 °C. The systems are equilibrated for another 4 ns. All nucleic acid atoms remain immobile throughout the course of the subsequent MD simulation to keep the structure from deviating from that of the crystal. Electrostatic interactions are computed with particle mesh Ewald summations. A 1 fs time step is used, and snapshots are saved every 1 ps. The simulations enforce NVT (constant number of atoms, volume, and temperature) conditions, because NPT (constant number of atoms, pressure, and temperature) simulations do not work well when a large number of atoms remain immobile.

4.5.2 HyPred

HyPred proceeds by calculating a set of proximal radial distribution functions (pRDF), defined in Table 2-1, for ~ 400 different categories of atoms in proteins and in RNA molecules using all-atom, explicit solvent MD simulations (17). This set of ~400 atom types distinguishes between atom and amino acid type, as well as nucleic acid base. The pRDFs describe the hydration layer near the nucleic acid’s surface atoms and are used to predict the solvent electron density in a discrete representation without additional MD simulations. For the purposes of calculating solvent electron density, the electrons are approximated as being localized at the nuclei of the solvent atoms, consequently, each oxygen, hydrogen, sodium, magnesium, and chloride atom/ion has 8, 1, 10, 10, 18 electrons, respectively.

The value of, for example, the pRDF \(g_{\text{CH}}(r)\) is calculated by averaging the density of all the cubes assigned to atom type CH and lying between a distance \(r - \Delta r\) and \(r + \Delta r\) from the CH
to which they are assigned. Illustrating this process for the pRDF $g_{\text{CH}}(r)$ for the atom type CH yields

$$g_{\text{CH}}(r) = \frac{1}{N} \sum_{i}^{N} \rho_i$$

where $\rho_i$ is the electron density of cube $C_i$, the summation is performed over all cubes that are assigned to H atoms of type CH, the cubes lie at a distance between $r - \Delta r$ and $r + \Delta r$ from the scaled van der Waals surface of the RNA’s H atom, and $N$ is the number of cubes in the summation. This procedure provides $g_{\text{CH}}(r)$, the pRDF for CH, which can be obtained by discretizing the simulation box into cubes.

The first step in calculating the pRDFs is the construction of the scaled van der Waals surface by surrounding all solute atoms with spheres of radii, $r=0.53r_{\text{vdw}}$ where $r_{\text{vdw}}$ is the standard van der Waals radius. The pRDFs are calculated by partitioning the MD simulation box into $(0.5 \text{ Å})^3$ cubes that are assigned to the solute atom whose scaled van der Waals surface is the closest to that cube, and the separation distance is stored. The densities and separations of the cubes assigned to each particular atom type are then used to calculate the entire pRDF for that atom type.

The hydration shell is recreated as follows using the same grid by inverting the process used to generate the pRDFs. We first store the distance between each cube that lies outside of the solute and the solute atom with the nearest scaled van der Waals surface. The solvent density in each cube is then assigned the value of the pRDF associated with that solute atom type and the separation. The model and the distribution functions for a given RNA are validated
by their ability to reproduce the hydration layer from the MD simulations for that RNA as well as those for other proteins

When HyPred is used to calculate the hydration shell density of a protein or RNA molecule, simulation data for that particular protein or RNA molecule are omitted, so the pRDF is evaluated as an average of pRDFs constructed from data for the other nucleic acids and proteins. Since the appearance of the publication originally describing HyPred (17), several improvements have been made to HyPred. The present work employs an enhanced procedure for calculating the pRDFs beyond those obtained from the original HyPred method (17). The first enhancement accounts for hydrogen bonding between solute and solvent. An improved description of hydrogen bonding is introduced by specifying the dependence of the pRDFs on the angle, termed the θ angle, formed by the vector originating at the center of the cube and ending at its assigned atom and the vector originating at the atom bonded to the atom assigned to the cube and ending at the atom assigned to the cube (Figure 3-5). If more than one atom is bonded to the solute atom nearest to the cube, we use the first bonded atom listed in a psf file. The bin size is 15°, and cubes are also categorized according to the atom type of the second closest atom in space, whereupon the atom type is specified by the elemental type of the second nearest heavy atom (e.g., S, O, N, or C) or, if the second nearest neighbor is a hydrogen atom, according to the heavy atom to which the H atom is bonded (e.g., CH, NH, OH). Cubes are also categorized according to whether they are near a convex or concave region of the protein surface. The list of atom types is expanded to include nucleic acid atom types. The RNA pRDFs are not constructed using any protein data.
4.5.3 Quantifying the quality of HyPred predictions

Three metrics are used to assess the ability of HyPred to reproduce the density maps obtained from the MD simulation. The first is the R factor which is defined by,

\[
R = \frac{\sum_i^N |\rho_{o,i} - \rho_i|}{\sum_i^N |\rho_{o,i} + \rho_i|}
\]

where \(\rho_{o,i}\) is the average solvent density in cube \(i\) as calculated from the MD simulation, \(\rho_i\) is the reconstructed density for that cube, and the summation runs over cubes that lie within 8 Å of the nucleic acid. The second error measure is the RMSD between the two densities (46),

\[
RMSD = \sqrt{\frac{\sum_i^N (\rho_{o,i} - \rho_i)^2}{\sum_i^N \rho_{o,i}}}
\]

The RMSD weighs more heavily the presence of regions with a large disparity between the reconstruction and the MD simulation than the real space R factor. Because the R factor and the RMSD strongly depend upon the extent of the bulk solution that is included in the calculation, we introduce another measure that is not as strongly dependent upon the amount of bulk solvent included in the reconstruction, provided that the simulation is long enough that bulk solvent density fluctuations, i.e., “noise”, is low. This third measure \(R^*\) is defined as

\[
R^* = \frac{\sum_i^N |\rho_{o,i} - \rho_i|}{\sum_i^N |\rho_{o,i} + \rho_i - 2\rho_3|}
\]
where $\rho_s$ is the bulk solvent density. The bulk solvent should not affect $R^*$ significantly because far from the nucleic acid, both $\rho_{o,i}$ and $\rho_i$ equal $\rho_s$, and the contribution of the bulk solvent to the numerator and denominator should each vanish.
5. HyPred based electrostatic calculations

5.1 Abstract

Electrostatics plays an essential role in many biological processes, but calculations of electrostatic solute-solvent interactions are computationally intensive. The HyPred theory for predicting the solvent density near the surfaces of proteins and nucleic acids is extended to predict the charge density near the surfaces of proteins and nucleic acids, thereby enabling a computationally quick calculation of the solute/solvent electrostatic interaction energy of hydration. A set of universal proximal radial electrostatic distribution functions (pREDFs) are defined and calculated for a series of atom types in proteins and RNA molecules using all-atom, explicit solvent molecular dynamics simulations for each of four ionic conditions. The charge density distribution is predicted for every protein (RNA) molecule using the average pREDFs of the other proteins (RNA) molecules, and the pREDFs are used to compute the solute/solvent electrostatic interaction energies. The predicted electrostatic energies of hydration from HyPred are compared to data from MD simulations and from Poisson-Boltzmann calculations performed using DelPhi for four different ionic conditions.

5.2 Introduction

The electrostatic energy of hydration is essential for protein and nucleic acid binding, molecular recognition(105), folding(105-107), and enzymatic activity(108). The study of electrostatics enters into descriptions of solvation energies, catalysis, redox potentials, and acid-base equilibria(109). While MD simulations can be used to calculate the electrostatic
energy and free energy of hydration\cite{110-112}, the simulations are computationally expensive. Thus, implicit solvent models, such as the Poisson-Boltzmann (PB)\cite{113}, Langevin-Debye (LD) (1925) and generalized Born (GB) models, \cite{114} have been developed at the expense of sometimes inaccurate approximations \cite{115}. DelPhi \cite{116} is a fast Poisson-Boltzmann equation solver that is commonly used to evaluate electrostatic free energies and provide insights into the molecular interactions and functions of RNA molecules \cite{117}. Pettitt et al. have introduced proximal electric radial distribution functions (pREDFs) as a means to predict the charge distribution around deca-alanine and the electrostatic free energy of hydration\cite{110}. Proximal radial distribution functions (pRDFs), the density analogs of pREDFs, are universal to all globular proteins \cite{6, 17, 118}. Here we present some extensions to Pettitt’s method that involve a more detailed atomistic description and a dependence upon the second nearest neighbor, orientation, and geometry of the solute molecule as in our HyPred theory \cite{17, 54}. The accuracy of the theory is demonstrated by applications to the more difficult case of highly charged RNA molecules and RNA binding proteins for a range of ionic conditions. We proceed by performing all-atom explicit solvent MD simulations for RNA molecules and proteins. The electrostatic energies of hydration are computed from the MD simulations. In order to analyze the electrostatic interactions, the simulation boxes are discretized into $0.4\ \text{Å}^3$ cubes, and the charge density in each cube is averaged over the simulation to produce charge density maps. The electrostatic energy of hydration is computed from these charge density maps to assess the error induced by discretizing the charge densities into cubes. Next, the charge density maps are used to calculate sets of universal (salt concentration dependent) pREDFs for water and ions, and the pREDFs are used to predict the charge density around proteins and RNA molecules.
without additional MD simulations. Comparisons of the charge densities determined from
HyPred and the MD simulations provide a stringent test of the accuracy of HyPred. Finally,
further comparisons of hydration energies from DelPhi enable assessing the accuracy all the
computational methods for four different ionic conditions.

5.3 Methods

5.3.1 MD simulations

All-atom explicit solvent MD simulations, employing NAMD(94) and the AMBER99(95)
force field, are performed for the RNA and protein portions of 1A1T(96), 1G70(99), 1HJI(100),
1S03(102), and 2A9X(103), separately. Four simulations are performed for each RNA molecule
and protein, each with a different concentration of ions, namely, 20 mM NaCl, 145 mM NaCl,
145 mM NaCl and 10 mM MgCl₂, and 1 M NaCl. The number of ions in each simulation is set by
the autoionize plugin of VMD, which first neutralizes the charges on the solute atoms and then
adds the specified concentration of ions. The number of each ion type and water molecule in
each simulation is listed in table 5-1. Each type of calculation reported here is repeated for each
of the four different ionic concentrations separately. Protonation is set by assuming standard
states for pH = 7. All simulations have box sizes whose dimensions extend 15 Å beyond solute
molecule on all sides. The density of the bulk solvent is determined in each case from all water
molecules situated at distances exceeding 10 Å from the solute. The electrons are
approximated as being located at the nuclei, and if the computed electron density of the bulk
water is either less than 0.333 e⁻ Å⁻³ or greater than 0.335 e⁻ Å⁻³, the simulation box size is
adjusted to produce the correct bulk density.
Table 5-1. The number of water molecules and ions in each simulation as well as the volume in Å³.

<table>
<thead>
<tr>
<th>Pdb</th>
<th>RNA Concentration</th>
<th>Water</th>
<th>Na</th>
<th>Cl</th>
<th>Mg</th>
<th>Volume</th>
<th>Protein Water</th>
<th>Na</th>
<th>Cl</th>
<th>Mg</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20 mM NaCl</td>
<td>18258</td>
<td>21</td>
<td>2</td>
<td>0</td>
<td>188622</td>
<td>22323</td>
<td>3</td>
<td>14</td>
<td>0</td>
<td>232197</td>
</tr>
<tr>
<td>1A1T</td>
<td>145 mM NaCl</td>
<td>18168</td>
<td>36</td>
<td>17</td>
<td>0</td>
<td>188622</td>
<td>22221</td>
<td>20</td>
<td>31</td>
<td>0</td>
<td>232197</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl 10 mM MgCl₂</td>
<td>18159</td>
<td>36</td>
<td>19</td>
<td>1</td>
<td>188622</td>
<td>22212</td>
<td>20</td>
<td>33</td>
<td>1</td>
<td>232197</td>
</tr>
<tr>
<td></td>
<td>1 M NaCl</td>
<td>17586</td>
<td>133</td>
<td>114</td>
<td>0</td>
<td>188622</td>
<td>21507</td>
<td>139</td>
<td>150</td>
<td>0</td>
<td>232197</td>
</tr>
<tr>
<td></td>
<td>20 mM NaCl</td>
<td>22836</td>
<td>34</td>
<td>3</td>
<td>0</td>
<td>240582</td>
<td>10809</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>108909</td>
</tr>
<tr>
<td>1G70</td>
<td>145 mM NaCl</td>
<td>22728</td>
<td>52</td>
<td>21</td>
<td>0</td>
<td>240582</td>
<td>10755</td>
<td>10</td>
<td>15</td>
<td>0</td>
<td>108909</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl 10 mM MgCl₂</td>
<td>22719</td>
<td>52</td>
<td>23</td>
<td>1</td>
<td>240582</td>
<td>10746</td>
<td>10</td>
<td>17</td>
<td>1</td>
<td>108909</td>
</tr>
<tr>
<td></td>
<td>1 M NaCl</td>
<td>21996</td>
<td>174</td>
<td>143</td>
<td>0</td>
<td>240582</td>
<td>10407</td>
<td>68</td>
<td>73</td>
<td>0</td>
<td>108909</td>
</tr>
<tr>
<td></td>
<td>20 mM NaCl</td>
<td>15141</td>
<td>16</td>
<td>2</td>
<td>0</td>
<td>154944</td>
<td>15243</td>
<td>2</td>
<td>8</td>
<td>0</td>
<td>155424</td>
</tr>
<tr>
<td>1HJI</td>
<td>145 mM NaCl</td>
<td>15069</td>
<td>28</td>
<td>14</td>
<td>0</td>
<td>154944</td>
<td>15171</td>
<td>14</td>
<td>20</td>
<td>0</td>
<td>155424</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl 10 mM MgCl₂</td>
<td>15060</td>
<td>28</td>
<td>16</td>
<td>1</td>
<td>154944</td>
<td>15162</td>
<td>14</td>
<td>22</td>
<td>1</td>
<td>155424</td>
</tr>
<tr>
<td></td>
<td>1 M NaCl</td>
<td>14583</td>
<td>109</td>
<td>95</td>
<td>0</td>
<td>154944</td>
<td>14685</td>
<td>95</td>
<td>101</td>
<td>0</td>
<td>155424</td>
</tr>
<tr>
<td></td>
<td>20 mM NaCl</td>
<td>36855</td>
<td>51</td>
<td>5</td>
<td>0</td>
<td>387585</td>
<td>32073</td>
<td>4</td>
<td>9</td>
<td>0</td>
<td>342385</td>
</tr>
<tr>
<td>1S03</td>
<td>145 mM NaCl</td>
<td>36687</td>
<td>79</td>
<td>33</td>
<td>0</td>
<td>387585</td>
<td>31923</td>
<td>29</td>
<td>34</td>
<td>0</td>
<td>342385</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl 10 mM MgCl₂</td>
<td>36669</td>
<td>79</td>
<td>37</td>
<td>2</td>
<td>387585</td>
<td>31905</td>
<td>29</td>
<td>38</td>
<td>2</td>
<td>342385</td>
</tr>
<tr>
<td></td>
<td>1 M NaCl</td>
<td>35499</td>
<td>277</td>
<td>231</td>
<td>0</td>
<td>387585</td>
<td>30897</td>
<td>200</td>
<td>205</td>
<td>0</td>
<td>342385</td>
</tr>
<tr>
<td></td>
<td>20 mM NaCl</td>
<td>21708</td>
<td>30</td>
<td>3</td>
<td>0</td>
<td>226207</td>
<td>10539</td>
<td>1</td>
<td>8</td>
<td>0</td>
<td>105978</td>
</tr>
<tr>
<td>2A9X</td>
<td>145 mM NaCl</td>
<td>21606</td>
<td>47</td>
<td>20</td>
<td>0</td>
<td>226207</td>
<td>10485</td>
<td>10</td>
<td>17</td>
<td>0</td>
<td>105978</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl 10 mM MgCl₂</td>
<td>21597</td>
<td>47</td>
<td>22</td>
<td>1</td>
<td>226207</td>
<td>10476</td>
<td>10</td>
<td>19</td>
<td>1</td>
<td>105978</td>
</tr>
<tr>
<td></td>
<td>1 M NaCl</td>
<td>20910</td>
<td>163</td>
<td>136</td>
<td>0</td>
<td>226207</td>
<td>10149</td>
<td>66</td>
<td>73</td>
<td>0</td>
<td>105978</td>
</tr>
</tbody>
</table>

Energy minimization and equilibration proceed in several stages. The solvent is first energy minimized for 200000 steps; then, with all atoms in the nucleic acids immobilized, the temperature is initially set at 10 K and then raised by 10 K every 100 ps until reaching the final temperature of 4 °C. The systems are equilibrated for another 4 ns. All solute atoms remain
immobile throughout the course of the subsequent MD simulation to keep the structure from deviating from that of the crystal. Electrostatic interactions are computed with particle mesh Ewald summations. A 1 fs time step is used, and snapshots are saved every 1 ps. The simulations enforce NVT (constant number of atoms, volume, and temperature) conditions, because NPT (constant number of atoms, pressure, and temperature) simulations do not work well when a large number of atoms remain immobile.

5.3.2 HyPred

HyPred proceeds by calculating a set of proximal radial distribution functions (pREDFs), defined in Table 2-1, for ~400 different categories of atoms in proteins and in RNA molecules using all-atom, explicit solvent MD simulations (17). This set of ~400 atom types distinguishes between atom and amino acid type, as well as nucleic acid base. The pREDFs describe the charge density distribution near the solute’s surface atoms and are used to predict the solvent charge density in a discrete representation. We use the charges from the protein structure file (psf) for the solute atoms. The first step in calculating the pRDFs is constructing the HyPred surface by surrounding all solute atoms with spheres with radii \( r = a r_{vdw} + b |q| \), where \( r_{vdw} \) is the van der Waals radius of the atom, \( q \) is the charge of the atom and \( a \) and \( b \) are optimized parameters. For RNA molecules \( a = 1.2 \) and \( b = 1.6 \text{ Å}^{-1} \), while for proteins \( a = 1.8 \) and \( b = 3 \text{ Å}^{-1} \). A cube is accessible to solvent if at least one of its corners can be encompassed by a sphere with a 1.4 Å radius which does not overlap with the van der Waals surface of the solute. The pREDFs are calculated by partitioning the MD simulation box into \((0.4 \text{ Å})^3\) cubes which are assigned to the solute atom whose HyPred surface is the closest to that cube, and the separation is stored. The average of the net charge densities within the cubes and the separations of the cubes form
their assigned particular atom type are then used to calculate the entire pREDf for that atom type. The charge density distribution in the hydration shell is recreated as follows using the same grid by inverting the process used to generate the pREDFs. We first store the distance between each cube that lies outside of the solute and the solute atom with the nearest scaled van der Waals surface. The solvent charge density in each cube is then assigned the value of the pREDf associated with that solute atom type and the separation. The model and the distribution functions for a given solute molecule are validated by their ability to reproduce the charge distribution and electrostatic energy from the MD simulations for that solute molecule as well as those for other proteins.

When HyPred is used to calculate the charge density of the solvent surrounding a protein or RNA molecule, simulation data for that particular protein or RNA molecule are omitted, so the pREDFs are evaluated as an average of pREDFs constructed from data for the other nucleic acids and proteins. Since the appearance of the publication describing HyPred(17), several improvements have been made to HyPred. The present work employs an enhanced procedure for calculating the pREDFs beyond those obtained from the original HyPred method (17). The first enhancement accounts for hydrogen bonding between solute and solvent. An improved description of hydrogen bonding is introduced by specifying the dependence of the pREDFs on the angle, termed the θ angle, between the vector connecting the center of the cube and the solute atom that is closest to the cube and the vector from the atom which bonded to the solute atom that is closest to the cube and the solute atom that is closest to the cube (Figure 3-5). If more than one atom is bonded to the solute atom nearest to the cube, we use the first bonded atom listed in a psf file. The bin size for θ is 15°, and cubes
are also categorized according to the atom type of the second closest atom in space, whereupon the atom type is specified by the elemental type of the second nearest heavy atom (e.g., S, O, N, or C) or, if the second nearest neighbor is a hydrogen atom, according to the heavy atom to which the H atom is bonded (e.g., CH, NH, OH). Cubes are also categorized according to whether they are near a convex or concave region of the protein surface. The average net charge density in each cube is also predicted by averaging the partial charge in each cube throughout the MD simulation in a straightforward extension of the procedure used for the electron density.

5.3.3 Electrostatic energy calculations

The protein or RNA/solvent electrostatic interaction energy is first evaluated from the MD simulations using,

$$E_{\text{elec}} = \sum_i^{N} \sum_j^{M} \frac{q_i q_j}{4\pi\varepsilon_o r_{ij}}$$

where $q_i$ is the charge on solute atom $i$, $q_j$ is the charge on solvent atom $j$, $r_{ij}$ is the distance between atoms $i$ and $j$, $N$ is the number of solute atoms, and $M$ is the number of solvent atoms. The electrostatic energy is calculated for every frame of the MD simulation using equation 1 and the individual electrostatic energies of all of the frames in the MD trajectory are averaged together. Next, the MD simulation box is partitioned into cubes, and the average charge density throughout the MD simulation is calculated in each cube. In order to test the accuracy of discretization of the solvent, equation 1 is then computed for the grid approximation by using the charges in each cube instead of the partial charges of the solvent atoms. Only cubes within
10 Å of the protein surface are included because the charge density is nearly zero further than
10 Å from the solute surface and the electrostatic potential is inversely proportional to the
distance. Next HyPred is used to predict the charge densities of the cubes in the hydration shell,
and the electrostatic energy is calculated from the solute/cube system. The protein or
RNA/solvent electrostatic interaction energy is also calculated with the Poisson-Boltzmann
method for comparison with the other calculations.

5.3.4 Nonlinear Poisson-Boltzmann calculations

For each of the proteins and nucleic acids studied the nonlinear Poisson-Boltzmann equation is
solved using DelPhi and $E_{\text{solute/solvent}}$ is evaluated using the energy(ion) option. The dielectric
constant of the solvent is set to 80. Since the solute atoms are fixed during the MD simulation
the motions of the solute atoms due not contribute to the dielectric constant, and since the
force field is not polarizable, polarizability does not contribute to the dielectric constant either.
Thus, the dielectric constant of the solute is set to 1. The van der Waals radii and charges on the
solute atoms are the same as the ones used to perform the MD simulations.

5.4 Results

5.4.1 Solute/solvent electrostatic interaction energies from MD simulations

Table 5-2 summarizes the electrostatic interaction energies $E_{\text{RNA/sol}}$, $E_{\text{RNA/water}}$, and
$E_{\text{RNA/ions}}$ calculated using equation 1 from MD simulations for 5 RNA molecules. As expected,
higher ionic concentrations generally lead to more favorable interactions $E_{\text{RNA/sol}}$. However,
perhaps surprisingly, the $E_{\text{RNA/sol}}$ experience only small changes as the ionic conditions are
varied, despite the significant contribution of the ions. Even at 20 mM NaCl, the lowest
concentration of NaCl for which we perform MD simulations, \( E_{\text{RNA/ions}} \) exceeds the \( E_{\text{RNA/water}} \) for all RNA molecules simulated here. The \( E_{\text{RNA/sol}} \) remain nearly constant because as the ionic concentration increases, the enhanced favorability of \( E_{\text{RNA/ions}} \) is largely offset by a decrease in the favorability of \( E_{\text{RNA/water}} \). For example, when the concentration of NaCl is increased from 20 mM to 1 M, \( E_{\text{RNA/ions}} \) grows from -3354 kcal mol\(^{-1}\) to -4117 kcal mol\(^{-1}\), an increase of 763 kcal mol\(^{-1}\), while the favorability of \( E_{\text{RNA/water}} \) decreases from -2517 kcal mol\(^{-1}\) to -1770 kcal mol\(^{-1}\), a decrease of 748 kcal mol\(^{-1}\). Thus, the total change of \( E_{\text{RNA/sol}} \) is 15 kcal mol\(^{-1}\) (only 0.26%) but is large compared to \( kT \).

The influence of ionic concentration on the properties of a solute molecule, such as binding to a ligand, is governed by the double difference of free energies, \( \Delta \Delta G \). For example, to determine the binding affinity of the RNA 1A1T to a protein as a function of ionic concentration, \( E_{\text{RNA/sol}} \) and several other energies associated with the 1A1T-ligand complex must be known as a function of ionic concentration. Thus, a change \( \Delta G \) of 15 kcal mol\(^{-1}\) in electrostatic energy must be considered in the context of the \( \Delta \Delta G \) of interest. Nevertheless, the \( \Delta G = 15 \) kcal mol\(^{-1}\) change in \( E_{\text{RNA/sol}} \) provides an order of magnitude estimate on the influence of the ions.

The minimal difference in total electrostatic energy \( E_{\text{RNA/sol}} \) with salt concentration exemplifies the remarkable ability of water to reorient and thereby counteract the influences of the charges. While this insensitivity may at first appear surprising, it is a consequence of the high dielectric constant of water. Two ions in water do not feel each other’s presence as greatly as two ions in vacuum because of the reorientation of the water molecules. Similarly, water screens the interaction between the ions and the charged atoms in the solutes. Ions situated
near energetically favorable locations of the RNA molecule displace water molecules, which would otherwise interact favorably with the RNA molecule.

We stress that the $E_{RNA/sol}$ is only a part of the total free energy of hydration because it omits the change in the solvent/solvent interaction energy and the entropy of ordering the solvent around the RNA, both of which are included in the free energy of hydration. A significant amount of free energy is required to separate solvent molecules in order to create a cavity capable of accommodating the RNA molecule. Another factor contributing to the free energy of hydration is the entropy due to the loss of hydrogen-bond exchange possibilities in the hydration shell of proteins and nucleic acids. Berry et al.(119) estimate the entropy of due to the reduction in the number of possible hydrogen-bonds near a solute with $\Delta s = k_B \ln \left[ \frac{m}{m-f} \right]$, where $m$ is the number of hydrogen bonds possible in water and $f$ is the number of hydrogen bonds lost by waters near the solute. If $m=4$, $f=1$ and the temperature is 298 K, than $\Delta s T$ is approximately 0.2 kcal mol$^{-1}$ of water in the hydration shell of a solute molecule. If a solute molecule is approximated as being a sphere with a 20 Å radius and 3 Å thick hydration shell than $\Delta s T$ due to the reduction in the number of hydrogen bonds in the hydration shell is approximately 90 kcal mol$^{-1}$ of solute.

When the NaCl concentration increases 50 fold from 20 mM to 1 M, the change in $E_{RNA/ions}$ does not exceed 32% for any of the RNA molecules studied. Thus, $E_{RNA/ions}$ already is approaching a plateau at 20 mM NaCl because the few cations in the simulation box readily find favorable positions proximal to the negatively charged groups of the RNA. Once the RNA is neutralized by these ions, the remaining cations are free to roam throughout the box, consistent with Manning counterion condensation theory(120). The 50 fold increase in ionic
concentration only refers to the ions beyond which are needed to neutralize the system. Thus, for example the number of sodium ions in the 2A9X simulations only increases by a factor of 5.4 when the NaCl concentration is increased from 20 mM to 1 M. Table 5-1 summarizes the actual number of ions used in each simulation.

The computed $E_{RNA/sol}$ are largely independent of ionic concentration for three reasons. Firstly, the change in $E_{RNA/water}$ with ionic concentration largely balances the change in the $E_{RNA/ions}$. Secondly, the first few ions in the solution screen the charges on the RNA molecule. Lastly, the water molecules also participate in screening the charges on the RNA molecule. The error bars for the $E_{RNA/sol}$ in Table 5-2 are smaller than the error bars for either $E_{RNA/water}$ or $E_{RNA/ions}$ because $E_{RNA/water}$ and $E_{RNA/ions}$ are anti-correlated from frame to frame in the MD simulations. Thus, the standard deviation of the sum of $E_{RNA/water}$ and $E_{RNA/ions}$ is less than the standard deviations of either the $E_{RNA/water}$ or $E_{RNA/ions}$.

Table 5-2. Solute/solvent electrostatic interaction energies of RNA molecules calculated from MD simulations.

<table>
<thead>
<tr>
<th>RNA</th>
<th>Ionic Condition</th>
<th>$E_{RNA/water}$ kcal mol$^{-1}$</th>
<th>$E_{RNA/ions}$ kcal mol$^{-1}$</th>
<th>$E_{RNA/sol}$ kcal mol$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A1T RNA</td>
<td>20 mM NaCl</td>
<td>$-2517 \pm 4$</td>
<td>$-3354 \pm 4$</td>
<td>$-5871 \pm 1$</td>
</tr>
<tr>
<td>1A1T RNA</td>
<td>145 mM NaCl</td>
<td>$-2201 \pm 5$</td>
<td>$-3678 \pm 5$</td>
<td>$-5879 \pm 1$</td>
</tr>
<tr>
<td>1G70 RNA</td>
<td>145 mM NaCl</td>
<td>$-2213 \pm 4$</td>
<td>$-3667 \pm 4$</td>
<td>$-5881 \pm 1$</td>
</tr>
<tr>
<td>1G70 RNA</td>
<td>10 mM MgCl$_2$</td>
<td>$-2213 \pm 4$</td>
<td>$-3667 \pm 4$</td>
<td>$-5881 \pm 1$</td>
</tr>
<tr>
<td>1G70 RNA</td>
<td>1 M NaCl</td>
<td>$-1770 \pm 6$</td>
<td>$-4117 \pm 6$</td>
<td>$-5887 \pm 1$</td>
</tr>
<tr>
<td>1G70 RNA</td>
<td>20 mM NaCl</td>
<td>$-3843 \pm 8$</td>
<td>$-7877 \pm 8$</td>
<td>$-11720 \pm 2$</td>
</tr>
<tr>
<td>1G70 RNA</td>
<td>145 mM NaCl</td>
<td>$-3529 \pm 8$</td>
<td>$-8187 \pm 8$</td>
<td>$-11716 \pm 2$</td>
</tr>
<tr>
<td>1G70 RNA</td>
<td>145 mM NaCl</td>
<td>$-3403 \pm 9$</td>
<td>$-8330 \pm 9$</td>
<td>$-11732 \pm 2$</td>
</tr>
<tr>
<td>10 mM MgCl₂</td>
<td>1 M NaCl</td>
<td>-2699 ± 8</td>
<td>-9040 ± 8</td>
<td>-11739 ± 2</td>
</tr>
<tr>
<td>-------------</td>
<td>----------</td>
<td>-----------</td>
<td>-----------</td>
<td>------------</td>
</tr>
<tr>
<td>20 mM NaCl</td>
<td>-1769 ± 4</td>
<td>-1883 ± 4</td>
<td>-3653 ± 1</td>
<td></td>
</tr>
<tr>
<td>145 mM NaCl</td>
<td>-1546 ± 4</td>
<td>-2110 ± 4</td>
<td>-3656 ± 1</td>
<td></td>
</tr>
<tr>
<td>145 mM NaCl</td>
<td>-1485 ± 4</td>
<td>-2175 ± 4</td>
<td>-3660 ± 1</td>
<td></td>
</tr>
<tr>
<td>10 mM MgCl₂</td>
<td>-1287 ± 4</td>
<td>-2372 ± 4</td>
<td>-3658 ± 1</td>
<td></td>
</tr>
</tbody>
</table>

**1HJI RNA**

<table>
<thead>
<tr>
<th>20 mM NaCl</th>
<th>-5889 ± 13</th>
<th>-16510 ± 13</th>
<th>-22398 ± 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>145 mM NaCl</td>
<td>-5446 ± 16</td>
<td>-16964 ± 16</td>
<td>-22411 ± 5</td>
</tr>
<tr>
<td>145 mM NaCl</td>
<td>-5140 ± 12</td>
<td>-17290 ± 12</td>
<td>-22430 ± 3</td>
</tr>
<tr>
<td>10 mM MgCl₂</td>
<td>-4267 ± 14</td>
<td>-18160 ± 13</td>
<td>-22427 ± 4</td>
</tr>
</tbody>
</table>

**1S03 RNA**

<table>
<thead>
<tr>
<th>20 mM NaCl</th>
<th>-3723 ± 9</th>
<th>-5863 ± 9</th>
<th>-9587 ± 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>145 mM NaCl</td>
<td>-3177 ± 8</td>
<td>-6420 ± 8</td>
<td>-9597 ± 3</td>
</tr>
<tr>
<td>145 mM NaCl</td>
<td>-3160 ± 6</td>
<td>-6437 ± 6</td>
<td>-9597 ± 3</td>
</tr>
<tr>
<td>10 mM MgCl₂</td>
<td>-2484 ± 9</td>
<td>-7128 ± 9</td>
<td>-9612 ± 3</td>
</tr>
</tbody>
</table>

**2A9X RNA**

<table>
<thead>
<tr>
<th>20 mM NaCl</th>
<th>-3723 ± 9</th>
<th>-5863 ± 9</th>
<th>-9587 ± 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>145 mM NaCl</td>
<td>-3177 ± 8</td>
<td>-6420 ± 8</td>
<td>-9597 ± 3</td>
</tr>
<tr>
<td>145 mM NaCl</td>
<td>-3160 ± 6</td>
<td>-6437 ± 6</td>
<td>-9597 ± 3</td>
</tr>
<tr>
<td>10 mM MgCl₂</td>
<td>-2484 ± 9</td>
<td>-7128 ± 9</td>
<td>-9612 ± 3</td>
</tr>
</tbody>
</table>

**Table 5-2 (continued)**

The $E_{prot/sol}$ exhibited in Table 5-3 are even less dependent on ionic conditions than are the $E_{RNA/sol}$, with the exception of the protein portions of the energies for 1A1T and 1S03. Many of the values $E_{prot/sol}$ for the 20 mM and 1 M NaCl solutions are equal within statistical error. The $E_{prot/sol}$ are smaller than the $E_{RNA/sol}$ because the protein atoms on average have less partial charge than the RNA atoms (see Table 5-4). The $E_{prot/ions}$ are less favorable than the $E_{prot/water}$ at 20 mM NaCl, while the opposite emerges for the RNA molecules because of the lower charges on the proteins. Hence, ions are less concentrated near the protein than near the RNA molecule. Moreover, since the water density is close to its maximum around both RNA
molecules and proteins, the difference between $E_{\text{prot/wat}}$ and $E_{\text{RNA/wat}}$ is less than the difference between $E_{\text{prot/ions}}$ and $E_{\text{RNA/ions}}$. The tendency of $|E_{\text{prot/sol}}|$ to be far less than $|E_{\text{RNA/sol}}|$ is a consequence of the greater charges on the RNA molecules.

Our method for accurately predicting the $E_{\text{solute/solvent}}$ in Tables 5-2 and 5-3 begins by computing solvent charge density maps (defined below) from the MD simulations as a function of ionic conditions. Next, we extract sets of pREDFs from the charge density maps and assess the influence of the ions and water on the pREDFs. The utility of the pREDFs is first tested by using them to reproduce the charge densities from the MD simulations and are compared with predictions from DelPhi.

Table 5-3. Solute/solvent electrostatic interaction energies of proteins calculated from MD simulations. Continued on next page.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Ionic Condition</th>
<th>$E_{\text{prot/water}}$ kcal mol$^{-1}$</th>
<th>$E_{\text{prot/ions}}$ kcal mol$^{-1}$</th>
<th>$E_{\text{prot/sol}}$ kcal mol$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A1T Protein</td>
<td>20 mM NaCl</td>
<td>-1750 ± 3</td>
<td>-1033 ± 3</td>
<td>-2783 ± 2</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl</td>
<td>-1610 ± 4</td>
<td>-1180 ± 3</td>
<td>-2790 ± 2</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl</td>
<td>-1585 ± 3</td>
<td>-1208 ± 3</td>
<td>-2793 ± 2</td>
</tr>
<tr>
<td></td>
<td>10 mM MgCl$_2$</td>
<td>-1311 ± 4</td>
<td>-1494 ± 4</td>
<td>-2805 ± 2</td>
</tr>
<tr>
<td>1G70 Protein</td>
<td>20 mM NaCl</td>
<td>-653 ± 2</td>
<td>-232 ± 1</td>
<td>-885 ± 1</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl</td>
<td>-592 ± 2</td>
<td>-292 ± 2</td>
<td>-884 ± 1</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl</td>
<td>-549 ± 3</td>
<td>-337 ± 2</td>
<td>-886 ± 1</td>
</tr>
<tr>
<td></td>
<td>10 mM MgCl$_2$</td>
<td>-465 ± 2</td>
<td>-420 ± 2</td>
<td>-886 ± 1</td>
</tr>
<tr>
<td>1HJ1 Protein</td>
<td>20 mM NaCl</td>
<td>-1004 ± 3</td>
<td>-359 ± 2</td>
<td>-1363 ± 1</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl</td>
<td>-937 ± 2</td>
<td>-424 ± 2</td>
<td>1362 ± 1</td>
</tr>
</tbody>
</table>
### Table 5-4. The charges on the protein and RNA molecules used in the simulations.

<table>
<thead>
<tr>
<th>Solute</th>
<th>Charge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A1T RNA</td>
<td>-19</td>
</tr>
<tr>
<td>1G70 RNA</td>
<td>-31</td>
</tr>
<tr>
<td>1HJI RNA</td>
<td>-14</td>
</tr>
<tr>
<td>1S03 RNA</td>
<td>-46</td>
</tr>
<tr>
<td>2A9X RNA</td>
<td>-27</td>
</tr>
<tr>
<td>1A1T Protein</td>
<td>11</td>
</tr>
<tr>
<td>1G70 Protein</td>
<td>5</td>
</tr>
<tr>
<td>1HJI Protein</td>
<td>6</td>
</tr>
<tr>
<td>1S03 Protein</td>
<td>5</td>
</tr>
<tr>
<td>2A9X Protein</td>
<td>7</td>
</tr>
</tbody>
</table>
5.4.2 Charge density maps

Charge density maps are generated from all-atom explicit solvent MD simulations for 5 proteins and RNA molecules by discretizing the MD simulation boxes with \((0.4 \text{ Å})^3\) cubes and then calculating the average net charge in each cube throughout the simulations. Figures 5-1a and 5-1b display cross sections of the charge density maps from the MD simulation of the separate RNA and protein portions of 2A9X. Many regions of high positive density (blue) are near the RNA due to waters orienting such that their hydrogen atoms near the negative charges of the RNA and to the presence of sodium ions. Similarly, water oxygen atoms are found near positively charged hydrogen atoms. Regions of high positive charge density are followed by regions of high negative charge density (red), a pattern imposed in part because of the alignment of multiple water dipoles. The charge density drops to approximately zero (white) within a few Ås of the surface of the RNA molecule.
The calculated cross sections of the charge density around the protein exhibit far fewer regions of high positive and negative density (Figure 5-1b), as expected. The net charge of the protein, displayed in Figure 5-1b is 7, and thus regions of high negative charge density appear close to the protein due to the proximal oxygen atoms of water and chloride ions. The regions of high negative charge density are followed by regions of high positive charge density, again because the negatively charged oxygen atoms are bonded to positively charged hydrogen atoms.
5.4.3 Dependence of charge density maps on ionic conditions

Figure 5-2 exhibits nearly identical charge density cross sections from the MD simulations of the RNA portion of 2A9X at four different ionic conditions (20 mM NaCl, 145 mM NaCl, 145 mM NaCl and 10 mM MgCl$_2$, and 1 M NaCl), while a similar behavior emerges in Figure 5-3 for the isolated protein. Very similar cross sections of the charge density map of 2A9X at 20 mM and 1 M NaCl again appear in Figures 5-4a,b as well as for the contribution of water to the charge density maps at 20 mM and 1M NaCl as presented in Figures 5-4 c,d. Regions of high positive or negative charge density are absent from the cross sections of the contribution of NaCl to the charge density at 20 mM NaCl, and only 4 regions of high charge density appear at 1M, implying that water provides the dominant contribution of charge density to the cross sections in Figures 5-4a, b. The sodium ions are more localized than the water molecules and produce a few small regions of high positive or negative charge density, which nevertheless contribute significantly to $E_{RNA/sol}$. Figures 5-5 e displays the regions in which the absolute value of the charge density contributed by sodium and chloride ions exceeds 0.5 qÅ$^{-3}$ in the simulation of the RNA portion of 2A9X at 20 mM NaCl. Figure 5-5 f contains two unexpected region of high negative charge density in the simulation of the RNA portion of 2A9X at 1 M NaCl that appear due to chloride ions near the negatively charged phosphate backbone. Their presence is due to the regions of high positive charge density from sodium ions and hydrogen atoms of water surrounding the phosphate backbone. The similarity of the distribution of charge densities due to the ions at 20 mM and 1 M NaCl indicates that the positions of the ions have converged, in the sense that another simulation of equal length would not likely produce a different charge density distribution. The charge density distribution
at 1 M NaCl only contains a few regions of high positive charge density due to sodium ions that are lacking in the charge density distribution at 20 mM NaCl, supporting the earlier statement that the influence of ions is already approaching a plateau at low ionic concentrations.

**Figure 5-2: Influence of ionic condition on charge density distribution around RNA.** The charge density distribution in a cross section of the RNA portion of 2A9X for four different ionic conditions.
Figure 5-3: Influence of ionic condition on charge density distribution around proteins. The charge density distribution in a cross section of the protein portion of 2A9X for four different ionic conditions.

The cross sections displayed in Figure 5-4 might lead to the belief that ions do not contribute significantly to $E_{RNA/sol}$. However, as discussed above, ions provide the major contribution to $E_{RNA/sol}$ because the dipole moment of water implies that the contributions from the hydrogen and oxygen atoms of water largely offset each other. Whenever the hydrogen atoms of water produce a region of high positive charge density near the solute, the oxygen atoms of water produce a region of high negative charge density appears further away. The pattern of alternative regions of high positive and negative charge density is absent from the charge density distribution of the ions. Instead, isolated regions of high positive charge from
sodium ions are evident at 20 mM NaCl but are mostly unaccompanied by regions of high negative charge density due to chloride ions. Since the separately mobile sodium and chloride ions do not locally counteract the effects of each other while the highly correlated oxygen and hydrogen atoms of water counterbalance each other, the contribution of the ions to $E_{RNA/sol}$ exceeds that of water.

Figure 5-4: The separate contributions of water and ions to the total charge density distribution at 20 mM and 1 M NaCl. (A and B) The total charge density distribution in the hydration shell. (C and D) The contribution of water to the charge density distribution. (E and F) The contribution of ions to the charge density distribution.
Figure 5-5: Three dimensional distribution of high positive and negative charge densities. Only regions for which the absolute value of the charge density is greater than 0.5 $\text{qÅ}^{-3}$ are displayed. (A and B) High charge density regions due to solution. (C and D) High charge density regions due to water. (E and F) High charge density regions due to sodium and chloride ions.

5.4.4 Charge pREDFs

A general method for calculating the charge density around arbitrary protein and RNA molecules begins with the extraction from the MD simulations of sets of universal pREDFs for the four different ionic conditions for each type of atom in protein and RNA molecules. The HyPred theory uses these pREDFs to calculate the charge densities surrounding any arbitrary protein or RNA molecule without further MD simulations. To calculate the pREDFs the HyPred surface is constructed by surrounding all solute atoms with spheres with radii $r = a_r \text{vdw} + b|q|$, where
where \( r_{vdw} \) is the van der Waals radius of the atom, \( q \) is the charge of the atom and \( a \) and \( b \) are optimized parameters. For RNA molecules \( a=1.2 \) and \( b=1.6 \, \text{Å}^{-1} \), while for proteins \( a=1.8 \) and \( b=3 \, \text{Å}^{-1} \). The solvent is discretized into cubes of dimensions \((0.4 \, \text{Å})^3\), and then every cube is assigned to the solute atom with the closest scaled van der Waals surface. The distances between each of the cubes and their closest solute atom's surface are then employed to calculate the pREDFs for the solute’s surface atoms by grouping cubes according to the atom type of the nearest solute atom and the distance to the nearest solute atom and by finding the average charge density as a function of the distance from the nearest solute atom and of the atom type of the nearest solute atom. Further refinement involves inclusion of the dependence of the pREDFs upon the second nearest neighbor, the complement of the angle formed by the line connecting the cube and the nearest solute atom and the line connecting the nearest solute atom and the atom to which it is bonded (Figure 3-5), and the concavity of the region in which the cube is located.

Figure 5-6 displays pREDFs for the CH and O atom types. Since hydrogen atoms bonded to carbon are not highly charged, the hydrogen atoms bonded to carbon only weakly attract ions and align the dipoles of water. Thus, only small net charge densities appear around hydrogen atoms bonded to carbon. Although hydrogen atoms carry slightly positive charges, the charge density right next to hydrogen atoms bonded to carbon is slightly positive because the smaller van der Waals radius of the hydrogen atom of water, relative to oxygen, enables the hydrogen to get closer to the solute atoms than an oxygen atom of water. Following the initial region with slightly positive density, the pREDF of CH becomes relatively large and negative because of the oxygen atoms of water, and the negative charge density region is followed by
three additional layers of alternating positive and negative charge density from the ordering of water molecules in the second and third water layers. The pREDF for the O atom type is highly positive at short distances because of the attraction of hydrogen atoms of water and positively charged ions to the negatively charged solute oxygen atoms. A region of high negative charge density emerges from the oxygen atoms of water that are bonded to the hydrogen atoms proximal to the solute oxygen atoms. Following the region of negative charge density, a second more diffuse region of positive charge density is caused by the second hydrogen atom of water, hydrogen atoms of other water molecules, and sodium ions.

Figure 5-6: Charge density pREDFs for the CH and O atom types at 20 mM and 1 M NaCl. Note the y-axis scale differ in the left and right panel.

5.4.5 Dependence of pREDFs on Ionic Conditions

The charge density maps are nearly independent of ionic conditions, and the pREDFs for the atom types CH and O at 20 mM NaCl and 1 M NaCl change very little with ionic conditions (Figure 5-6). When the NaCl concentration is increased from 20 mM to 1 M, the separate changes in the contributions of water and the ions to the pREDF for the O atom type in Figure
5-7 are constructive for some distances and destructive for some others. The change with ion concentration in the contribution of water to the charge density distribution resembles in some ways to the change in the contribution of NaCl to the charge density distribution, except that the change in the contribution of water to the pREDF is shifted to larger distances and contains an initial negative charge density peak and two additional negative charge density peaks, whereas the change in the contribution of NaCl to the pREDF lacks negative peaks. The absolute value of the greatest peak in the change to the contribution of the charge density is about 0.004 \( q \text{Å}^{-3} \), which is small in comparison to the 0.09 \( q \text{Å}^{-3} \) peak in the pREDF of the O atom type.

![Graph of charge density distribution](image)

**Figure 5-7: The difference in the pREDFs for the O atom type at 1 M and 20 mM NaCl.** The blue curve is the total change in the pREDF, the black curve is the change due to ions, and the red curve is the change due to water.
In addition to the charge densities maps, the molar concentration maps of each solvent atom type are generated in the same way that charge density maps are obtained, except that instead of averaging the net charge density the average number density of each solvent atom type in each cube is found separately and converted to molarity. The molar concentration maps are used to calculate proximal radial distribution functions (pRDFs), in exactly the same manner as the pREDFs are calculated from the charge density maps. Figure 5-8 displays the pRDFs for the sodium and chloride ions that are associated with the CH and O atom types of RNA. The pRDFs in Figure 5-8 exhibit a tendency towards saturation of the sodium ion concentration near the solute atoms when the bulk concentrations of NaCl are 145 mM and 1 M. The peak in the sodium ion pRDF of the CH atom type is 1.2 M at 145 mM NaCl and 2.6 M at 1 M NaCl. Thus, a nearly 7 fold increase in the bulk sodium ion concentration only increases the peak in the pRDF by little more than a factor of 2 because the first few ions near a group screen the charges on the RNA molecule. The sodium ions near solute oxygen atoms are even closer to saturation, as evidenced by the peak in the sodium pRDF of the O atom type of 5.4 M at 145 mM NaCl and 8.4 at 1 M NaCl because the density of sodium ions is high to begin with, and thus the sodium ions repel additional sodium ions more strongly than for the CH atom type. The peak in the sodium ion pRDF of the CH atom type appears at 4 Å where the chloride ion pRDF is 0.1 M when the NaCl concentration is 145 mM, but the peak grows to 0.6 M when the NaCl concentration is increased to 1 M. The chloride ion concentration at the peak increases 6 fold, while the sodium ion concentration only doubles because the increased concentration of sodium ions attracts chloride ions. The chloride ion pRDF of the O atom type vanishes for distances less than 4 Å.
even when the NaCl concentration is 1 M due to the strong repulsion of chloride ions by the oxygen atoms of RNA.

\[
\begin{align*}
\text{RNA 145 mM} & \quad \text{RNA 145 mM} \\
\text{RNA 1.0 M} & \quad \text{RNA 1.0 M}
\end{align*}
\]

Figure 5-8: The sodium and chloride ion pRDFs for the CH and O atom types from the 145 mM and 1 M NaCl MD simulations.

5.4.6 Comparison between protein and RNA ion pRDFs

Figure 5-9 illustrates the pRDFs (in M) of sodium and chloride ions for the CH and O atom types of RNAs and proteins. Obviously, because of the signs of the charges on the molecules, the concentration of sodium ions always exceed the concentration of the chloride ions in the solvent surrounding RNA molecules, while the opposite is true for the positively charged (RNA binding) proteins studied. Thus, the concentration of chloride ions near the CH
atom type of proteins exceeds the concentration of sodium ions, while the opposite naturally appears for RNA molecules because the CH atoms only weakly influence the distribution of ions, whereas their neighbors strongly affect the distribution of ions. The sodium ion pRDFs of the O atom type exhibit very large peaks because the sodium ions are attracted to the negatively charged oxygen atoms.

Figure 5-9: Comparison of protein and RNA sodium and chloride ion pRDFs at 145 mM and 1 M NaCl.

5.4.7 Solute/solvent electrostatic interaction energies from charge density maps

Tables 5-5 and 5-6 present the $E_{\text{solute/solvent}}$ calculated from the discretized charge density maps obtained from MD simulations using equation 1. The $E_{\text{solute/solvent}}$ deduced from
the discretized charge density distributions provide an upper bound for the accuracy of any method for calculating $E_{\text{solute/solvent}}$ using a lattice of cubes with predicted charge densities. Generally, only small errors are introduced by discretizing the average net charge density into $(0.4 \text{ Å})^3$ cubes. The largest discretization error here is 0.36%, while the largest error in absolute terms is 34 kcal mol$^{-1}$. The calculated $E_{\text{solute/solvent}}$ from the density maps tend to be more dependent upon the ionic concentration than the $E_{\text{solute/solvent}}$ from the MD simulations. For example, the difference in $E_{\text{RNA/sol}}$ at 20 mM NaCl and at 1 M NaCl calculated from the MD simulation is 15 kcal mol$^{-1}$ for the RNA portion of 1A1T, while the difference in $E_{\text{RNA/sol}}$ at 20 mM NaCl and 1 M NaCl calculated from the charge density maps is 24 kcal mol$^{-1}$.

5.4.8 Charge density maps predicted by HyPred

Hydration shell charge densities of the individual protein RNA molecules are reconstructed by using the pREDFs from the MD simulations and reversing the procedure followed for calculating the pREDFs. When calculating the charge density distribution for a protein or RNA molecule, the pREDFs of that molecule are omitted. Instead HyPred uses the pREDFs of the other protein or RNA molecules.

A cross section of the HyPred predicted solvent charge density map of the RNA portion of 2A9X, displayed in Figure 5-1b, exhibits many of the same features as the charge density map calculated from the MD simulation presented in Figure 5-1a. For example, a region of high positive charge density followed by two regions of high negative charge appears near the (C31)O2P atom in both the charge density map obtained from the MD simulation and the
HyPred prediction, and many other similarities are present near the (C8)O2’, (U10)O1P, (U10)O2’, (C15)O2’, and (C23)O2’ atoms.

The accuracy of the cross section of the charge density map predicted by HyPred for the protein portion of 2A9X, presented in Figure 5-1d, is comparable to that of the HyPred prediction for the RNA portion of 2A9X. HyPred correctly predicts regions of high negative density followed by regions of high positive density near Arg H\_ε and the C-terminal oxygen atoms and a region of high positive charge density appearing near the carbonyl oxygen atom of Arg5.

5.4.9 Solute/solvent electrostatic interaction energies calculated by HyPred

The E\_solute/solvent are calculated from the charge density maps predicted by HyPred, using equation 1. The HyPred calculated E\_solute/solvent of the proteins and RNA molecules studied here are presented in Tables 5-5,5-6. The best prediction made by HyPred is for the protein portion of 1A1T in 20 mM NaCl solution, in which case HyPred predicts an E\_solute/solvent of -2776 kcal mol\(^{-1}\), compared with -2783 kcal mol\(^{-1}\) obtained from the MD simulation. Thus, the relative error of the HyPred prediction of E\_solute/solvent compared to the MD simulation is only 0.25% or 7 kcal mol\(^{-1}\). The HyPred predictions for the other ionic conditions tend not to be as accurate as the predictions for the proteins and RNA molecules in 20 mM NaCl solution, and the changes in E\_solute/solvent with respect to ionic condition predicted by HyPred are inaccurate. The origins of HyPred’s inability to reproduce the dependence of E\_solute/solvent on ionic conditions are still under investigation. The worst prediction made by HyPred for E\_solute/solvent of a protein in 20 mM NaCl solution is for 1S03, in which case the E\_solute/solvent predicted by HyPred -2037 kcal mol\(^{-1}\) while the E\_solute/solvent provided by the MD simulation is -2381 kcal mol\(^{-1}\). Thus, the difference
between the predicted $E_{\text{solute/solvent}}$ and $E_{\text{solute/solvent}}$ obtained from the MD simulation is 344 kcal mol$^{-1}$ or about 14%. On average the HyPred predictions for the RNA molecules tend to be more accurate than the predictions for the proteins. The average relative error of the HyPred predictions compared to the MD simulations for the RNA molecules at 20 mM NaCl is 2.3% while the average error for the proteins is 5.2%. The best HyPred prediction for the RNA molecules in 20 mM NaCl solution in terms of relative error in comparison to the MD simulation is for 1S03. The $E_{\text{solute/solvent}}$ predicted for the RNA portion of 1S03 is -22234 kcal mol$^{-1}$, while the $E_{\text{solute/solvent}}$ obtained from the MD simulation is -22398 kcal mol$^{-1}$. While the HyPred estimate for the $E_{\text{solute/solvent}}$ of the RNA portion of 1S03 is 164 kcal mol$^{-1}$ too small the error is only 0.73% of the $E_{\text{solute/solvent}}$ calculated from the MD simulation. The worst HyPred prediction for the RNA molecules in 20 mM NaCl solution is for 2A9X. The $E_{\text{solute/solvent}}$ predicted by HyPred for the RNA portion of 2A9X is -9194 kcal mol$^{-1}$, while the $E_{\text{solute/solvent}}$ calculated from the MD simulation is -9587 kcal mol$^{-1}$, or about a 4.1% error.

**Table 5-5: Comparison of the solute/solvent electrostatic interaction energies for RNA molecules calculated from MD simulations, charge density maps obtained from MD simulations, HyPred, and nonlinear Poisson-Boltzmann equation.** Continued on next page.

<table>
<thead>
<tr>
<th>Solute</th>
<th>Ionic Condition</th>
<th>MD simulation kcal mol$^{-1}$</th>
<th>Density Map kcal mol$^{-1}$</th>
<th>HyPred kcal mol$^{-1}$</th>
<th>DelPhi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A1T RNA</td>
<td>20 mM NaCl</td>
<td>-5871 ± 1</td>
<td>-5865</td>
<td>-5818</td>
<td>-5353</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl</td>
<td>-5879 ± 1</td>
<td>-5878</td>
<td>-6137</td>
<td>-5362</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl</td>
<td>-5881 ± 1</td>
<td>-5886</td>
<td>-5797</td>
<td>-5365</td>
</tr>
<tr>
<td></td>
<td>10 mM MgCl$_2$</td>
<td>-5881 ± 1</td>
<td>-5886</td>
<td>-5797</td>
<td>-5365</td>
</tr>
</tbody>
</table>
Table 5-5 (continued)

Table 5-6: Comparison of the solute/solvent electrostatic interaction energies for proteins calculated from MD simulations, charge density maps obtained from MD simulations, HyPred, and nonlinear Poisson-Boltzmann equation. Continued on next page.

<table>
<thead>
<tr>
<th>Solute</th>
<th>Ionic Condition</th>
<th>MD simulation</th>
<th>Density Map kcal mol(^{-1})</th>
<th>HyPred kcal mol(^{-1})</th>
<th>DelPhi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1G70 RNA</td>
<td>1 M NaCl</td>
<td>-5887 ± 1</td>
<td>-5889</td>
<td>-6389</td>
<td>-5370</td>
</tr>
<tr>
<td></td>
<td>20 mM NaCl</td>
<td>-11720 ± 2</td>
<td>-11719</td>
<td>-11915</td>
<td>-10959</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl</td>
<td>-11716 ± 2</td>
<td>-11715</td>
<td>-12495</td>
<td>-10976</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl</td>
<td>-11732 ± 2</td>
<td>-11755</td>
<td>-11433</td>
<td>-10980</td>
</tr>
<tr>
<td></td>
<td>10 mM MgCl(_2)</td>
<td>-11739 ± 2</td>
<td>-11773</td>
<td>-12507</td>
<td>-10989</td>
</tr>
<tr>
<td>1HJI NA</td>
<td>20 mM NaCl</td>
<td>-3653 ± 1</td>
<td>-3641</td>
<td>-3496</td>
<td>-3301</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl</td>
<td>-3656 ± 1</td>
<td>-3643</td>
<td>-3629</td>
<td>-3307</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl</td>
<td>-3660 ± 1</td>
<td>-3656</td>
<td>-3374</td>
<td>-3309</td>
</tr>
<tr>
<td></td>
<td>10 mM MgCl(_2)</td>
<td>-3660 ± 1</td>
<td>-3656</td>
<td>-3374</td>
<td>-3309</td>
</tr>
<tr>
<td></td>
<td>1 M NaCl</td>
<td>-3658 ± 1</td>
<td>-3662</td>
<td>-3695</td>
<td>-3312</td>
</tr>
<tr>
<td>1S03 RNA</td>
<td>20 mM NaCl</td>
<td>-22398 ± 3</td>
<td>-22385</td>
<td>-22234</td>
<td>-20954</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl</td>
<td>-22411 ± 5</td>
<td>-22394</td>
<td>-21955</td>
<td>-20957</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl</td>
<td>-22430 ± 3</td>
<td>-22419</td>
<td>-21292</td>
<td>Failed to converge</td>
</tr>
<tr>
<td></td>
<td>10 mM MgCl(_2)</td>
<td>-22427 ± 4</td>
<td>-22436</td>
<td>-22929</td>
<td>-20961</td>
</tr>
<tr>
<td>2A9X RNA</td>
<td>20 mM NaCl</td>
<td>-9587 ± 3</td>
<td>-9580</td>
<td>-9194</td>
<td>-8916</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl</td>
<td>-9597 ± 3</td>
<td>-9609</td>
<td>-9167</td>
<td>-8930</td>
</tr>
<tr>
<td></td>
<td>145 mM NaCl</td>
<td>-9597 ± 3</td>
<td>-9607</td>
<td>-8140</td>
<td>-8933</td>
</tr>
<tr>
<td></td>
<td>10 mM MgCl(_2)</td>
<td>-9612 ± 3</td>
<td>-9642</td>
<td>-9233</td>
<td>-8941</td>
</tr>
<tr>
<td>Protein</td>
<td>20 mM NaCl</td>
<td>145 mM NaCl</td>
<td>145 mM NaCl 10 mM MgCl₂</td>
<td>1 M NaCl</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>------------------</td>
<td>------------------</td>
<td>--------------------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>1A1T Protein</td>
<td>-2783 ± 2</td>
<td>-2790 ± 2</td>
<td>-2793 ± 2</td>
<td>-2805 ± 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-2769</td>
<td>-2787</td>
<td>-2786</td>
<td>-2799</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-2776</td>
<td>-2797</td>
<td>-3122</td>
<td>-3256</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-2751</td>
<td>-2756</td>
<td>-2756</td>
<td>-2761</td>
<td></td>
</tr>
<tr>
<td>1G70 Protein</td>
<td>-885 ± 1</td>
<td>-884 ± 1</td>
<td>-886 ± 1</td>
<td>-886 ± 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-888</td>
<td>-889</td>
<td>-881</td>
<td>-891</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-832</td>
<td>-750</td>
<td>-621</td>
<td>-748</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-907</td>
<td>-908</td>
<td>-910</td>
<td>-910</td>
<td></td>
</tr>
<tr>
<td>1HJI Protein</td>
<td>-1363 ± 1</td>
<td>-1362 ± 1</td>
<td>-1364 ± 1</td>
<td>-1364 ± 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1362</td>
<td>-1362</td>
<td>-1358</td>
<td>-1366</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1409</td>
<td>-1530</td>
<td>-1556</td>
<td>-1454</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1377</td>
<td>-1380</td>
<td>-1380</td>
<td>-1382</td>
<td></td>
</tr>
<tr>
<td>1S03 Protein</td>
<td>-2381 ± 2</td>
<td>-2377 ± 2</td>
<td>-2383 ± 2</td>
<td>-2397 ± 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-2380</td>
<td>-2379</td>
<td>-2386</td>
<td>-2401</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-2037</td>
<td>-2022</td>
<td>-2548</td>
<td>-2251</td>
<td></td>
</tr>
<tr>
<td>2A9X Protein</td>
<td>-1071 ± 1</td>
<td>-1071 ± 1</td>
<td>-1072 ± 1</td>
<td>-1071 ± 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1067</td>
<td>-1070</td>
<td>-1071</td>
<td>-1074</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1049</td>
<td>-1094</td>
<td>-1039</td>
<td>-1155</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1127</td>
<td>-1129</td>
<td>-1129</td>
<td>-1131</td>
<td></td>
</tr>
</tbody>
</table>

Table 5-6 (continued)
5.4.10 Comparison to DelPhi

For each of the proteins and nucleic acids studied DelPhi is used to solve the nonlinear Poisson-Boltzmann equation and evaluate $E_{\text{solute/solvent}}$. For most of the proteins, the DelPhi prediction of $E_{\text{solute/solvent}}$ is good. For 1A1T, 1G70, and 1HJI the percent errors for the 20 mM NaCl solutions are 1.1%, 2.5%, and 1.0% respectively while the percent errors are 5.2% and 16.7% for 2A9X and 1S03 respectively. HyPred predicts the $E_{\text{solute/solvent}}$ for the protein portions of 1A1T, 1S03 and 2A9X in 20 mM NaCl solution more accurately than DelPhi. For 1A1T HyPred is off by only 0.25% from the MD simulations while DelPhi is off by 1.1%. For 2A9X the relative error of the HyPred prediction in comparison to the MD simulation is 2.1% and DelPhi is off by 5.2%. For 1S03 HyPred is off by 14.4% and DelPhi is off by 16.7%. On average the relative error for HyPred predictions in comparison to the MD simulations are 5.2% and the error for the DelPhi predictions is 5.3%.

The DelPhi predictions for the RNA molecules are worse than for the proteins. The DelPhi predictions for the RNA molecules are consistently smaller than the values obtained from the MD simulation by between 6.4% and 9.6%, probably due to the highly charged nature of the RNA molecules. The average percent error of the DelPhi predictions for the RNA molecules is 7.7%, while the average percent error of the HyPred predictions is 2.3%. The DelPhi predictions for the change in $E_{\text{solute/solvent}}$ with ionic condition are inaccurate, though DelPhi outperforms HyPred in this regard. In the cases of the RNA portions of 1A1T and 2A9X the differences between $E_{\text{solute/solvent}}$ of the 20 mM and 1 M solutions obtained from the MD simulations are 15 and 25 kcal mol$^{-1}$ respectively while the predictions by DelPhi are 17 and 25 kcal mol$^{-1}$. However in other cases the value of the change in $E_{\text{solute/solvent}}$ with ionic condition obtained using DelPhi is not close to the
value calculated from the MD simulations. For example, DelPhi predicts that the difference in $E_{\text{solute/solvent}}$ between the RNA portion of 1S03 in 20 mM NaCl solution and 1 M NaCl solution is 29 kcal mol$^{-1}$, while the corresponding value from the MD simulations is only 7 kcal mol$^{-1}$. While HyPred is able to give estimates of $E_{\text{solute/solvent}}$ for all cases, DelPhi is unable to give a reasonable value for $E_{\text{solute/solvent}}$ of the RNA portion of 1S03 in a solution of 145 mM NaCl and 10 MgCl$_2$.

The agreement between DelPhi and the MD simulations could be improved by adjusting the parameters used to carry out the calculations. The values of $E_{\text{solute/solvent}}$ reported by DelPhi are sensitive to the van der Waals radii(111, 121). The dielectric constant of the solvent used in the DelPhi calculations should be the same as the dielectric constant of the water in the MD simulations, but the dielectric constant of water in the MD simulations has not yet been evaluated. The values of the van der Waals radii can be adjusted to optimize agreement with the MD simulations.

5.5 Conclusion

We present a model for the accurate and rapid calculation of the charge density around proteins and nucleic acids and the solute/solvent electrostatic interaction energy. Water is found to make a significant contribution to the charge density around proteins and nucleic acids. The charge density distribution around proteins and nucleic acids is more complicated than depicted by continuum models such as PB. Charge density distributions around proteins and nucleic acids have alternating heterogeneous regions of high positive and negative charge densities. The HyPred model largely maintains the accuracy of MD simulations, while being faster by orders of magnitude. HyPred based electrostatic calculations can still be improved. A few issues need to be addressed before this work can be published. First HyPred must be able
to predict $E_{\text{solute/solvent}}$ accurately at ionic conditions other than 20 mM NaCl. As more proteins and RNA molecules have been used to construct the pREDFs, the predicted $E_{\text{solute/solvent}}$ have improved. Enough protein and RNA simulations must be used to calculate $E_{\text{solute/solvent}}$ such that the results converge. The $E_{\text{solute/solvent}}$ predicted by HyPred are sensitive to the manner in which the solvent accessibility of the cubes are determined and HyPred can and should be improved in its ability to determine solvent accessibility. The DelPhi calculations also need to be improved. The dielectric constant of bulk water was set to 80, but the dielectric constant of the bulk water in the MD simulations is not necessarily 80. The dielectric constant of the bulk water in the MD simulations must be evaluated and utilized in the DelPhi calculations. The $E_{\text{solute/solvent}}$ obtained using DelPhi are also dependent upon the inputted van der Waals radii. The van der Waals radii used by DelPhi should be optimized. Future applications include the calculation of binding energies.
6. Conclusion

6.1 Summary

The HyPred theory for modeling water around proteins and nucleic acids is developed and applied to predicting X-ray crystallographic waters, SWAXS patterns, and solute/solvent electrostatic interactions energies. HyPred is formulated based on the assumption that the hydration of a solute molecule is determined by local properties of the solute molecule. The HyPred model is many orders of magnitude faster than MD simulations and has proven to be surprisingly accurate compared to MD simulations.

6.2 Future directions

Since hydration is important to many aspects of biology, multiple applications exist for HyPred, but only a few applications are discussed in this thesis. Solvation is influenced by many factors and HyPred takes only a few of these factors into account. Enabling HyPred to model more of the factors which influence solvation should improve the accuracy of HyPred. The following section briefly discusses some projects I have worked on, but have not completed, or would work on if I continued to study hydration.

6.2.1 Binding

The role of hydration in protein-RNA interactions is unclear. Work has been initiated to use HyPred in order to predict protein-RNA binding, but no firm conclusions have been reached yet. This work begins with all-atom explicit solvent molecular dynamics simulations for RNA molecules in complexes with different proteins and in isolation.
I proceed by superimposing the hydration shell of the free protein (RNA) onto the hydration shell of the complex. The portion of the free hydration shell of the protein (RNA) which has no corresponding region of nonzero density in the hydration shell of the protein-RNA complex is referred to as the portion of the protein (RNA) hydration shell which is buried upon binding. In order to study the differences between the hydration shell of the free protein (RNA) and the portion of the hydration shell of the free protein (RNA) which are buried upon binding, I calculate the solvation density as a function of distance from the protein (RNA) surface by combining all atom types into one atom type and evaluating pRDFs, termed general pRDFs, using this specification separately for the free proteins and RNA molecules and for the portions of the hydration shells which are buried upon binding. General pRDFs are only dependent upon the distance to the nuclei of the nearest solute atom and have no conditionals on the atom type of the nearest solute atom or any other factors. General pRDFs of the densities of the free hydration shells and the portions of the hydration shells that are buried upon binding are evaluated from MD simulations and are compared to those predicted by HyPred. Figure 6-1 displays the general pRDFs of the portions of the protein and RNA hydration shells which are buried upon binding and the general pRDFs of the free protein and RNA for 1A1T. The portion of the RNA hydration shell of 1A1T which is buried upon binding is less dense than the average RNA hydration shell, while the buried portion of the protein hydration shell is slightly less dense than the average protein hydration shell, leading to the hope that in all cases the portion of the hydration shell that is buried upon binding is less dense than the hydration shell of free proteins and RNAs.
However, this result is not universal. The general pRDF of the buried portions of the hydration shell of 1G70 is very similar to the general pRDF of the free protein and RNA. General pRDFs for the portions of the hydration shells of the free proteins and RNAs which are buried upon formation of the protein-RNA decoys using densities predicted by HyPred have not proven to be helpful in selecting the native protein-RNA complex from decoys generated by FTDock (122). However, continued work may yield useful conclusions. For example, while the densities of the portions of the hydration shells that are buried upon binding are not necessarily lower than those of the hydration shells as a whole, a correlation might exist between the change in the protein hydration shell upon burial and the change in the RNA hydration shell upon burial. Specific and non-specific protein-RNA binding pairs may have different hydration patterns, which can be used to distinguish between them.

Crystal structures often exhibit water molecules that are shared by proteins and nucleic acid molecules (83, 123) and that also appear in the crystal structures of the separate protein and nucleic molecules. Water molecules that are displaced during binding may also play a role in binding, motivating an investigation of the correspondence between the displaced protein and RNA water molecules. My goals are to describe the shared water molecules in MD simulations, predict those present in the complexes, and use this information, along with other characteristics of the hydration shell, to guide the selection of the native protein-RNA complex from a set of decoys. The hydration shells of the free protein and free RNA molecule are superimposed and the portion of the protein (RNA) hydration shell which overlaps with the hydration shell of the RNA (protein) and is within 3.5 Å of both the protein and RNA is referred to as the portion of the protein (RNA) hydration shell which intersects the RNA (protein).
hydration shell. I generate what I refer to as plots of the shared waters by extracting regions of high density from the portion of the hydration shell of the free protein (RNA) which intersects with the hydration shell of the free RNA (protein) when the hydration shells of the free protein and RNA are superimposed, superimposing these regions of high density, and plotting the number of such regions of high density originating from the hydration shell of the free RNA molecules which have such a region of high density originating in the hydration shell of the free protein within a radius of r is plotted as a function of r. Plots of the shared waters are generated from the density maps of the MD simulations and the hydration shell density maps predicted by HyPred.

In addition HyPred is used to model the hydration shells for a set of protein-RNA decoys, and plots of the shared waters are obtained for the decoys. We attempt to assess the differences between the hydration of the native protein-RNA complexes and the decoys by examining the plots of the shared waters of the native and decoy conformations.

No clear conclusions emerge from the results available. Plots of the shared waters are displayed in Figure 6-2, for density maps obtained from the MD simulation of the native complex, HyPred prediction for the native complex, and HyPred predictions for 10 decoys, generated by FTDock (122). No general conclusions can be drawn from Figure 6-2, since the results from the MD simulation and HyPred are similar to the decoys.
The general pRDFs of the hydration shell of free RNA portion of 1A1T (black) and the region of the hydration shell of the RNA portion of 1A1T, that is buried upon burial (red).

The general pRDFs of the hydration shell of free protein portion of 1A1T (black) and the region of the hydration shell of the protein portion of 1A1T, that is buried upon burial (red).

The general pRDFs of the hydration shell of free RNA portion of 1G70 (black) and the region of the hydration shell of the RNA portion of 1G70, that is buried upon burial (red).

The general pRDFs of the hydration shell of free protein portion of 1G70 (black) and the region of the hydration shell of the protein portion of 1G70, that is buried upon burial (red).

Figure 6-1: General pRDFs of the hydration shells of the separate protein and RNA portions of 1A1T and 1G70, along with general pRDFs of the regions of the hydration shells that are buried upon binding.

The methods introduced in the previous chapter to evaluate electrostatic energies of...
hydration may be used to predict the electrostatic free energy of binding, in a method similar to the one used by DelPhi to predict the electrostatic free energy of binding (124).

Figure 6-2: The number of RNA water molecules with a protein water molecule within a distance of \( r \). The thick black line is the results of the MD simulation. The thick red line is the HyPred calculation on the native. The thin lines are HyPred calculations on a set of 10 decoys.

6.2.2 Orientation of water molecules

The ability to predict the orientations of water molecules in the hydration shells of macromolecules would have many applications such as a greater understanding of electrostatics and entropy of hydration. Much of the entropy of hydration arises due to the ordering of water molecules around solutes. Density maps of discretized water orientations can
be generated from MD simulations, orientational pRDFs can be computed from the density maps, and this process can be reversed to generate predicted orientational density maps for proteins and nucleic acids for which MD simulations have not been performed. Figure 6-3 displays separate pRDFs for the CH and NH atom types for the cases in which the angle formed by the dipole moment of water and the vector originating at the proximal solute atom and ending at the oxygen atom of water is greater than 90° (red) and less than 90° (blue). The distribution of water dipoles in a cube combined with information on the local field can help with the calculation of dielectric constants. Thus, a three dimensional map of the dielectric could be generated from MD simulations, pRDFs of the dielectric could be obtained and HyPred could be used to predict the dielectric around proteins and RNA molecules.

**Figure 6-3: Orientational pRDFs:** pRDFs for the CH and NH atom types for when the angle formed by the dipole moment of water and the vector originating at the proximal solute atom and ending at the oxygen atom of water is greater than 90° (red) and less than 90° (blue).
6.2.3 Free energy of hydration

The free energy of hydration is critical in determining events such as protein folding and binding and can be estimated from the three dimensional distribution of water densities and orientations. Free energy calculations may be a particularly good application for HyPred since it takes an extremely long time for free energies obtained from MD simulations to converge. The potential ability of HyPred to estimate free energies of hydrations could be used to pick out native like protein structures from a set of decoys. The hydration entropy and free energy along the folding pathway of proteins could be studied with HyPred.

6.2.4 Residency times

HyPred gives a static picture of the hydration shell. The calculation of residency times would give a picture of the dynamics of water in the hydration shell. The dynamics of waters around proteins has a significant impact on the dynamics of proteins (5). The correlation between residency times and density could be studied using MD simulations and HyPred. A comparison of the residency times of waters around 1L2P and hydration shell density, displayed in Figure 6-4, indicates a correlation between residency times and density. However this is the only case for which I have computed residency times and I have not made a quantitative assessment of the correlation between residency times and density. Previous studies on the relationship between residency times and densities have found little correlation between the two (125, 126).

For some protein properties, residency times might be more important than density. Residency times of water molecules have been studied in the past (125, 126), but to my
knowledge nobody has attempted to predict residency times. HyPred could predict residency times in a manner similar to the method by which HyPred predicts the other solvation properties of solutes. Three dimensional distributions of residency times of water molecules could be calculated by discretizing an MD simulation box into cubes and averaging residency times of water molecules in each cube throughout the MD simulations. Residency pRDFs could be obtained from the three dimensional distribution of residency times, and the average residency times of water molecules could be predicted around proteins and nucleic acid molecules for which MD simulations have not been carried out.

Figure 6-4: Residency times compared with density

A) Cross section of residency times
B) Cross section of density

**Figure 6-4: Residency times compared with density** A) A cross section of 1L2P with long residency times in red and short residency times in blue. B) A cross section of 1L2P displaying the hydration shell density.
6.2.5 Hydrogen bonding

Hydrogen bonding is central to the hydrophobic effect and energetics of hydration. Much of the ordering of water molecules around hydrophobes results from the fact that fewer conformationations are available for water molecules which maintain a network of hydrogen bonds around hydrophobic groups. For example Berry et. al. estimate the reduction in entropy due to the reduced number of hydrogen bonds available to a water molecule near a hydrophobe (119). Calculating pRDFs of water-water hydrogen bond densities and the number of water-water hydrogen bonds per water molecule and using the water-water hydrogen bond pRDFs to predict hydrogen bond densities around proteins and nucleic acids could be fruitful.

6.2.6 Crystallographic refinement

A method similar, to the HyPred bases approach presented in chapter 3 for calculating SWAXS patterns, could be used to generate crystallographic scattering patterns, and refine crystal structures. Haddadian et al. have developed a method for automated real-space refinement of protein structures(127). HyPred could be combined with this algorithm.
6.2.7 Improving the accuracy of HyPred

The size of a spherical hydrophobic particle influences the density of water in its hydration shell. Proteins and nucleic acids are not perfect spheres and instead exhibit many bumps and crevices. The size and curvature of the local bumps and crevices of a solute molecule might be more important to solvation of the solute molecule than the overall size of the solute molecule. Though HyPred makes a binary assignment of cubes into concave or convex regions, HyPred does not quantify the local curvature. The predictions obtained from HyPred could be made more accurate by quantifying the local curvature of the solute. Local geometry, specified by concavity, has a smaller influence on the pRDFs than the second nearest neighbor or angular dependence, but this could be an artifact due to the fact that HyPred groups together many dissimilar types of local geometries.

Regions in which a water molecule can form multiple hydrogen bonds with a solute may be particularly dense and HyPred may underestimate the density of such regions. Taking into account the possibility of forming multiple hydrogen bonds could help improve the accuracy of HyPred. The influence of the third nearest neighbor could also be taken into account. Since adding more conditionals would reduce the amount of statistics per category, the third nearest neighbor would simply be classified as being either hydrophilic or hydrophobic. Alternatively instead of using the second nearest neighbor as a conditional the number of hydrophilic groups within a certain radius could be used as a conditional. Water molecules in the hydration shells of solutes are not only influenced by the solute, but also by other water molecules. Water-water distribution functions could be used to improve HyPred.
6.2.8 Effect of temperature and pressure

Due to an interesting temperature dependence of the hydrophobic effect, resulting from changes in the entropic and enthalpic cost of reordering water in the hydration shell (128) with temperature, proteins are unstable at both high and low temperatures. It is important to understand this temperature dependence and reproducing the temperature dependence of the hydrophobic effect would be a strong validation of the HyPred model. MD simulations could be carried out at a variety of temperatures and a set of pRDFs could be obtained for each temperature. A greater understanding of the temperature dependence on the hydrophobic effect could be gained by comparing the hydration around proteins from thermophiles, mesophiles, and psychrophiles. Pressure also has an effect on the hydration of proteins and nucleic acids. For example many proteins undergo pressure denaturation. Studying the impact of hydration on pressure denaturation at an atomic level using MD simulations would be interesting. A comparison of the hydration of proteins from piezophiles and mesophiles could be productive.

6.2.9 Different water models

Most of the work presented in this thesis is performed using the TIP3P water model, which does not entirely capture the tetrahedral structure of water as well as the TIP4P and TIP5P water models (129). Many of the interesting properties of water which arise from the tetrahedral structure of water may be missed by using the TIP3P water model. The hydration shells of BPTI obtained from MD simulations carried out with the TIP3P and TIP4P water models displayed in Figure 6-5 are similar. BPTI is only one example and the density from the TIP4P
simulation has not yet converged. Using the TIP5P water model may provide more insight into hydration, as it reproduces the tetrahedral geometry of water more accurately than the TIP3P or TIP4P water models. MD simulations performed with the SPC/E (130) model are in better agreement with experiments than MD simulation performed with the TIP3P model (131), so simulations carried out using the SPC/E water model may prove interesting. A polarizable water model (132) may also provide interesting and accurate results.

Figure 6-5: A comparison of the TIP3P and TIP4P water models.

6.2.10 Development of an implicit solvent model

All-atom explicit solvent simulations are extremely computationally expensive. The computational expense of MD simulations can be drastically reduced by eliminating the water and replacing the water by an implicit solvent model. Many implicit solvent models suffer from inaccuracies (133). If HyPred could be used to calculate free energies of hydration then the forces on each atom can be found by taking the derivative of the free energy of hydration with
respect to the atomic coordinates. However, using HyPred directly in an implicit solvent
simulation would be too slow. Nevertheless, HyPred might give insights into the interaction
between biomolecules and solvent that can be applied to the development of an implicit
solvent model. If free energies of hydration could be obtained from HyPred then structural
features could be correlated with the free energies obtained from HyPred. Many implicit
solvent models express the free energy of hydration as a weighted sum of the accessible
surface areas of different atom types(53). HyPred could provide better weights for the
accessible surface areas, or a better method of estimating free energies than using accessible
surface areas. In addition studying the dielectric constants provided by HyPred could produce a
better model of the dielectric constants around proteins and nucleic acids.
References


