COMPUTATIONAL STUDY OF WATER AND ION DISTRIBUTIONS AROUND BIOMOLECULES

BY HUNG TIEN NGUYEN

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Computational Biology and Molecular Biophysics

Written under the direction of

Prof. David A. Case

and approved by

____

____

____

____

New Brunswick, New Jersey

Jan, 2016

ABSTRACT OF THE DISSERTATION

Computational study of water and ion distributions around biomolecules

by Hung Tien Nguyen

Dissertation Director: Prof. David A. Case

Water and ions play a crucially important role in governing biomolecule structure, stability and function. Knowledge of how water molecules and ions distribute around proteins and nucleic acids at the molecular level has long been sought. Due to their highly mobile nature, the hydration water and ion cloud are very hard to probe with traditional experiment techniques such as X-ray crystallography, NMR or microscopy. Here we use a combination of computational approaches and X-ray scattering experiment to investigate the water and ion distribution around biomolecules.

In the first part, we describe a protocol to calculate X-ray scattering profiles from atomic models of macromolecules. We show that the quality of the Reference Interaction Site Model (RISM) hydration closely approaches those from explicit molecular dynamics simulation in terms of reproducing X-ray intensity signals. The intensity profiles (which involve no adjustable parameters) match experiment and molecular dynamics simulation up to wide angle for relatively rigid biomolecules. For nucleic acid structures, we demonstrate that an improvement in the intensity calculations could be made by using the conformational ensemble obtained from MD simulation rather than using a single diffraction structure.

In the second part, we extend the X-ray scattering theory and describe a novel analysis method to extract water and ion distribution from X-ray scattering experiment. The analysis complements recent experimental techniques, showing both numbers of excess solvent (water, ions) and aspects of their distributions around macromolecules. Comparisons between experimental and theoretically predicted distributions are made for molecular dynamics and RISM theory, showing that although the total X-ray patterns are very similar, the distributions from MD simulations are generally better than those from RISM. This illustrate the potential power of this analysis to guide the development of computational models of solvation.

Finally, we investigate a possible use of partial molar volume and number of excess solvent (extracted from X-ray experiment and from other direct measurements) as a guide to recalibrate force fields. The partial molar volume (and the number of excess solvent) can be conceptually divided into contributions of the solute excluded volume and hydration shell. While the former depends only on the solute topology and can be computed once the solute structure is known, the latter is more “interesting” and contains valuable information about solute-solvent interaction. We show that current protein force fields reproduce reasonably the hydration shell term although more works are needed to achieve better solute-solvent interaction balance. For nucleic acids, the solute-solvent interaction is strongly overestimated and a recalibration is needed. As a proof of concept, we reoptimize the non-bonded parameters for the phosphate groups in a DNA duplex and show that the predicted partial molar volume and the number of excess hydration water around the DNA approach the experimental value. Our parameters, however, currently cannot be used for dynamics study unless a complete refit of bonded parameters is carried out. Since the nucleic acid structure depends tightly on the solute-solvent interaction, we believe that such a misbalance should be corrected in the near future.

Acknowledgements

I feel extremely lucky to be able to finish this dissertation with the advice from my advisor and committee members, help from my friends and love from my family.

I would like to express my deepest gratitude to my supervisor, Prof. David Case, who has been supporting and providing me valuable advice during my PhD study. I thank him for his guidance and effort he put into training me to become a "toddler" scientist. He has been supportive and has always given me the freedom to pursue my own thoughts without objection. I appreciate his patience to correct my reports and manuscripts due to my limited writing skills. I will still remember his first email responding to me back when I was applying to Rutgers graduate program as a prospective student. For me, it is an honor to become his student and work in his lab.

I am very grateful to Prof. Lois Pollack for her scientific advice and many insightful discussions and suggestions. I would also like to thank Prof. Darrin York and Prof. Wilma Olson for valuable comments at all levels of my research. My research would not be possible without their helps.

I also wish to thank Dr. George Giambasu and Dr. Suzette Pabit who provide good suggestions for my research project. Many thanks go to Prof. Tyler Luchko and Dr. In Suk Joung who teach me the basis of RISM theory. I am thankful to Jesse Johnson for many stimulating chats. I also want to acknowledge other members of Case's group and York's group for their friendship.

I take this opportunity to express my gratitude to my beloved parents, grandparents and my sisters. They have always been supporting and encouraging me with their best wishes.

Finally, I would like to thank my wife and my son for their understanding, love, kindness and moral support. You two make this adventure such an amazing experience.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Water in biology

1.1.1 Role of water and hydration

Water is the most important substance on planet Earth as it is a crucial solvent for life to evolve and sustain. In living cells, water performs many important functions such as transporting, stabilizing, lubricating, reacting, partitioning ... and thus cannot be considered as a simple dilutent. It is widely accepted that biomolecules such as proteins and nucleic acids will not function as usual without the presence of liquid water. (There are some exceptions that the proteins could retain their function in nonaqueous solvent or at low hydration condition, [20, 21] but in both cases there are still some tightly bound water molecules remain.) The effect of water to life is extremely fine-tuned to a degree that even introducing heavy water is toxic to those processes.

There has been enormous effort to understand the role of water in molecular interaction, for some recent excellent reviews see [22–30]. There is abundant experimental and theoretical evidence that the water molecules adjacent to the protein (referred to as "hydration water") has structural and dynamical properties very different from the bulk water. However, ambiguities appear once a rigorous definition for hydration water is required. Since an experimental technique can only probe one aspect of the hydration water, different approaches thus lead to different operational definitions of hydration. [31]

It is worth noting that the thermodynamic and structural hydration is totally independent of dynamic hydration. For example, one dynamical aspect of hydration water is the residence time–the inverse of the first-order dissociation constant. A large residence time of a water molecule does not mean that water molecule has a stronger affinity towards the protein, nor does it contribute to the excess density of water at that location. It could simply come from the fact that this water molecule is stuck inside a deep cavity, and since the interior of proteins comprises mostly of hydrophobic residues, its interaction with the protein is not necessarily stronger than the interfacial water molecules. Thermodynamic and structural properties depend only on the local minima on the potential energy surface, while dynamics are controlled by the barrier height and the saddle points. There are lots of studies to investigate the effect of biomolecules onto dynamical aspect of hydration water or vice versa. The interested readers are referred to excellent works in the literature, for instance see reviews by Halle, [32] Bagchi, [23] Pal and Zewail [30] and Fogarty *et al.* [33] This dissertation devotes only to study of the structural feature of hydration water.

1.1.2 Structure of hydration water

Currently, there is no experimental method with enough spatial and temporal resolution to directly probe water molecules interacting with biomolecules. The current view of hydration water mostly comes from more or less model-dependent interpretation of the experimental data. X-ray crystallography is one of few methods successfully used to probed internal water [34, 35] (water molecules occupy cavities within the protein and are conserved as the amino acid sequence). Those water molecules are usually considered as the essential and integral part of the protein structure. However, detecting external water molecules with diffraction methods clearly requires more care. The hydration sites in the crystal are not necessarily the same as in solution since water molecules in the crystal usually meditate protein-protein interaction. Furthermore, the presence of salting-out agents, ions and contaminants in the crystal could further complicates the situation. In fact, when different structures of the same protein are aligned, only a few hydration sites are found to be conserved. [36]

X-ray and neutron scattering studies suggest that the density of the hydration layer around globular proteins is approximately 10% denser than the bulk. [37, 38] This value is later confirmed by MD simulation by Merzel and Smith. [39] Noteworthy, two-thirds of the observed density increase are merely caused by the protein surface existence. This

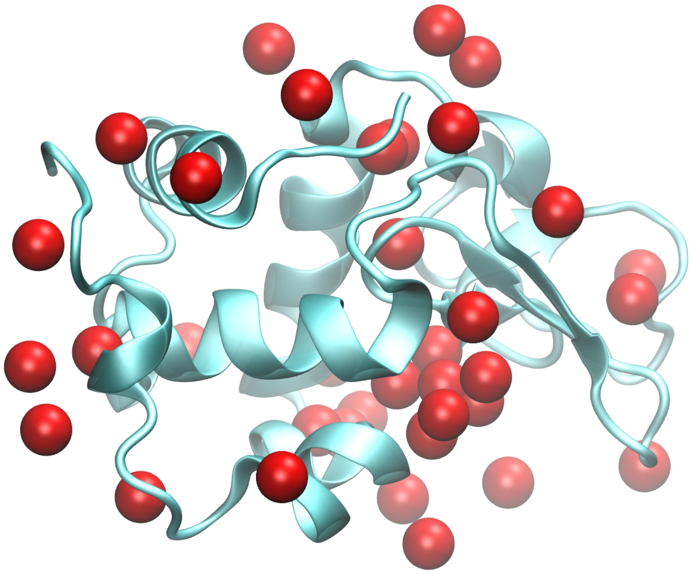


Figure 1.1: Crystallographic structure of hen egg-white lysozyme (PDB 6LYZ). Also shown in red spheres are water molecules revealed in the electron density maps (only water molecules that are within 3 Å from the protein are shown).

contribution would arise even if the water molecules were not perturbed by the presence of the protein. In their simulation, they also observed the shorten of water-water distance and increase of the coordination number of water in the hydration shell.

1.1.3 Theoretical efforts to study hydration structure

Pettitt and colleagues find that in general, a continuous distribution density of hydration water is more appropriate than an atomistic description. [27, 40] Such an approach roots back from the theory of liquid state in which the concept of pair distribution function is of central importance. [41] Using molecular dynamics simulation, they showed that although hundreds of hydration sites could be located for myoglobin, only half of them were occupied at any time. Noteworthy, there was almost no significant difference between the calculated free energies between conserved and nonconserved solvent sites. They concluded that any given set of ordered solvent molecules represents one of many possible configurations and description of solvation based on such one–or even many–is incomplete. To construct the solvent distribution density, they proposed the so-called proximal radial distribution functions (pRDF) of hydration water based on the assumption that the solvent structure around the solute is dominated by the local interaction. Additional contributions from farther atoms were assumed small and could be incorporated later as the perturbation to the distribution function.

Similar efforts include the work of Hummer *et al.* who used the potential of mean force (PMF) to construct the solvent density map. [42, 43] In their approach, the distribution of water around a biomolecule was approximated by the distribution of water around other water molecules, with the position of those water molecules being at the location of the electronegative atoms of the biomolecule (O, N, S), *i.e.* equating all electronegative atoms in the solute to water oxygen with respect to their effect on ordering water. Additional modification was incorporated to include steric factors and hydrophobic regions (*e.g.* non-polar atoms were treated as hard spheres ...) Other works including AquaSol, which represents the solvent as a collection of orientable dipoles with nonuniform concentration, and semi-explicit assemble (SEA), which uses precomputed properties of water solvation around simple molecules from explicit simulation to generate a combined solvation shells around an arbitrary macromolecule, exist in the literature. [44, 45] Recent studies that use more intricate treatment for water include those coming from the Integral Equation theory–Reference Interaction Site Model (RISM, which will be discussed more detailed in Chapter 2) and its relative "classical" density functional theory of solvation (which will not be considered in this dissertation, readers are referred to [46–49]).

1.1.4 Molecular dynamics simulation

Molecular dynamics simulation is one widely used tool to investigate the hydration water problem. Of course, simulation studies of hydration structure require both reasonably accurate water models and mature biomolecular force fields. Although there has been a vast number of water models proposed (ranging from quantum, polarizable, fixed-charge, coarse-grained models), [24, 50, 51] the steady appearance of new models in recent years indicates that there are still lots of room of improvement and that the current widely used water models probably need a serious reconsideration. [52–55] To make matters worse, biomolecular force fields are built on those water models and therefore, strictly speaking, a force field only should be used with its corresponding water model. Integrating a new water model into the simulation requires a systematic and serious validation which is tedious and usually takes years to decades.

In the simulation, water-protein interaction strength undoubtedly plays a crucial role governing the protein shape and fold. A change of water model used in the simulation has a significant effect onto protein stability and structure. [56–58] If the modeled water-protein interaction is too strong compared with the internal protein interaction (between residue-residue within the protein) then the protein tends to expand and unfold, maximizing its exposed surface with water. On the other hand, if this interaction is weak then the protein will remain compact. More and more evidence suggests that this is a very subtle balance, and even a very small change of water-protein interaction could lead to a significant difference of protein structures observed in the simulation (see below). Current biomolecular force fields have always been built to work with folded structures. In fact, one of the earliest criteria to validate a protein force field is the ability to maintain the protein near its crystal structure. This could inadvertently put more weight onto the internal protein interaction while underestimate the water-protein interaction.

Recent studies, while working with expanded protein conformations such as in protein folding and intrinsically disordered proteins, reveal that there is probably a misbalance between water-protein and protein-protein interactions. Best and coworkers report that unfolded or disordered states of proteins are too collapsed using the current force fields, indicating that proteins are poorly solvated and the non-specific protein-protein interaction appears to be too strong (compared with water-protein interaction). [59] The same conclusion is also drawn from the work of Piana *et al.* who find that the water models significantly underestimate the dispersion interaction. [55] By increasing the water interaction, they show that the disordered states of proteins are substantially more expanded and are generally in better agreement with the experiment. Other works by different groups have also reached similar conclusions. [60–62]

Such a misbalance for nucleic acid force fields also exists as illustrated by an interesting work of Chen and Garcia. In their paper, they show that it is necessary to adjust both the base-base and water-base interactions to reversibly fold small RNA tetraloops from their unfolded states. [63] The fact that the base-base stacking strength is increased while the water-base interaction is reduced in their approach illustrates that the water-nucleic acid interaction is probably too strong. In another relevant work, using a different water model which emphasizes the importance of electrostatics interaction, [53] Bergonzo and Cheatham show that the structure of tetranucleotides is remarkably improved in the simulation when comparing with NMR experiment. [64] Force fields for nucleic acids are under very active development, which focuses on the modification of internal structural aspects of nucleic acids via torsional angle parameters (see [65] and references therein). However, given the fact that the structure is influenced strongly by water interaction, a recalibration of water-nucleic acid interaction is another worth pursuing possibility.

1.2 Ion distribution

1.2.1 Introduction

Similar to water, ions are ubiquitous in living system and play important role in supporting biomolecule function, stability, dynamics and folding. Monovalent salts (Na^+, K^+, Cl^-) are crucially vital in regulating the homeostasis and electric potentials of cells. Sodium is the most abundant ion in human plasma and biological fluids. Transport of ions through membranes is responsible for sound, smell, sight, taste and touch we (humans) perceive daily. Ions are also known to actively participate in catalytic activity, and not just nonspecific ionic buffering agents. [66]

Due to its high charge, nucleic acids need counterions to maintain their shape and fold. [67–70] The interaction between nucleic acids could be changed dramatically, from repelling to attracting, merely by adjusting the ionic strength of the solution. [71–73] The counterions are therefore considered as an integral and essential part of the polyelectrolyte, with the name “ion cloud” or “ionic atmosphere”. [74–76] As a result, the dynamics and interaction between nucleic acids and their complexes with proteins cannot be fully understood without a reasonable description of the ionic atmosphere. From a theoretical perspective, interaction between counterions and nucleic acids is dominated by electrostatics, nevertheless there is also contribution from specific ion effect. This results two types of counterion in the ion cloud: a loosely associated sheath of ions surrounding the macromolecules (at long distance, due to the electrostatic interaction) and some tightly bound ions at specific location right on the solute surface (due to specific ion effect). Furthermore, the fluctuating nature of both the ion cloud and the polyelectrolyte makes understanding the ionic atmosphere challenging, both experimentally and theoretically.

1.2.2 Theoretical efforts to study the ion cloud

A natural treatment of the ion cloud is to ignore the specific effect and consider only the electrostatic interaction between ions and the polyelectrolyte. Most of the theoretical efforts relies on the Poisson–Boltzmann (PB) equation, which roots from pioneering works of Gouy–Chapman, Debye–Huckel, Onsager, Kirkwood, Tanford. As pointed out by Kirkwood, the basic approximation of the PB equation is the replacing of the potential of mean force by the mean electrostatic potential. [77] Short range interaction and ion-ion correlations are ignored in the PB aproach. (Some modifications later are added into the theory to account for finite size of ions and include the ionic correlations. [78]) Still, the theory is capable to make lots of successful predictions. [79–83]

One of the early theories that directly targets the ionic atmosphere around the nucleic acids is the counterion condensation theory proposed by Manning. [84,85] In the theory, the polyelectrolyte was assumed as an infinite wire (or later, cylinder) with a uniform charge density. The PB equation was then solved to obtain the ion distribution around the polyelectrolyte. Using this theory, Manning showed that counterions “condensed” near the polynucleotide, neutralizing a portion of the solute charge to bring its effective charge down to a critical value (for instance, the charge fractions of DNA neutralized by Na^+ and Mg^{2+} in aqueous solutions are, respectively, 76% and 88% of its total charge).

With the computational and algorithmic advance in the eighties and nineties of the last century, more complicated and accurate treatments of the systems are allowed. The solute eventually could be described with full atomic details, while solvent (including ions) is still treated at the continuum level. [86,87] The PB equation can be solved numerically to obtain a 3-dimensional electrostatic potential map. In practice of this method, Honig and coworkers showed that the deepest potentials around the B-DNA were located in the grooves rather than near the phosphate regions, which was in accordance with earlier quantum calculation. [88, 89] This partly explained why ions were found in both the major grooves (GC-rich regions) and in narrow minor grooves (A-tracts). [90–93]

The groove binding pattern of counterions is also observed in early MD simulations. [94–97] Na^+ is shown to bind in both grooves with the preferred binding site in the minor groove while K^+ mostly binds to the major groove and close to the center of B-DNA. Direct binding of ions to DNA bases are rarely observed during the simulation, but once it occurs, the ion could remain there for an extended period of time (on the order of tens nanoseconds). Binding to the phosphate group is more uniform. In all cases, there is always at least a water molecule bridging the ion and the nucleic acid. Due to the fluctuating nature of the ion cloud, however, there was always a question of convergence of the simulation during the last decade. Recent efforts, assisted by advances of computational algorithms and resources, expand the time scale of the simulation (up to hundreds nanosecond) and investigate ion competition towards nucleic acids. [9, 98–101]

1.2.3 Experimental studies

Compared with the theory side, quantitative experimental research of the ion atmosphere notably lags behind. The reason is because the ion atmosphere is mostly invisible to traditional structural biology techniques. X-ray crystallography can only detects strongly bound ions (mostly typically bound to specific sites on the biomolecules) which account for a small fraction of the ion cloud. [102, 103] Similarly, NMR relaxation approach only assesses the ions closely "attached" to the nucleic acids, but does not take into account the ions that are distant away. [90, 104] Methods such as ion-specific fluorescent dyes also have been used to count specific ions interacting with polynucleotides, but cannot determine the ion cloud as a whole. [105]

A recent experiment technique, buffer equilibration–atomic emission spectroscopy (BE–AES) or "ion counting", allows an assess of the full content of the ion cloud (see Fig. 1.2 for

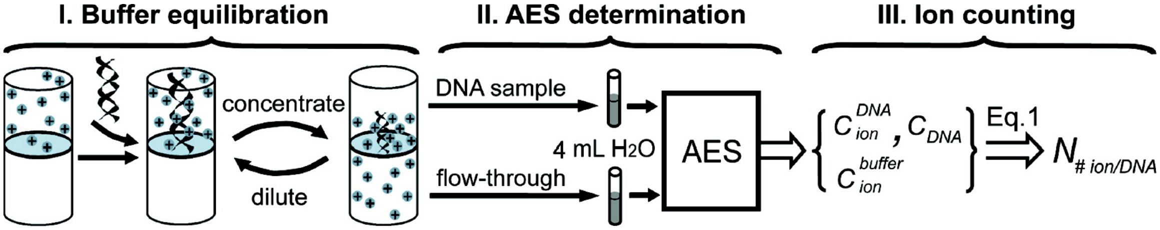


Figure 1.2: Scheme of the buffer equilibration–atomic emission spectroscopy (BE–AES), a.k.a. "ion counting" approach. Reprinted with permission from Bai *et al.* [7] Copyright 2007 American Chemical Society.

a schematic description of the experiment). [7] The buffer around the nucleic acid sample is first fully equilibrated. Then, by comparing the ion concentration (monitored by AES) between the DNA-contained sample with the buffer-only sample, the total number of excess ions present in the ion cloud could be counted. The experiment shows that in order to neutralize the DNA charge, not only counterions are attracted towards the polyelectrolyte but also at the same time the co-ions (which have the same charge as the DNA) are expelled far away. More importantly, the sum of the excess numbers of counterions and co-ions must match the total charge of the DNA, *i.e.* the system is totally neutralized. However, this kind of experiment does not provide any information about the shape of the ion atmosphere (how far the counterions distribute themselves around the solute).

A different method that can provide the number of excess ion is anomalous small-angle X-ray scattering (ASAXS). [106] The method will be discussed in more details in Chapter 4. Essentially the method employs the fact that the "effective number" of electrons in the ions could be changed by varying the X-ray beam energy near a critical value (the ion absorbance edge). This energy variation only affects the ions in the solution, and leaves those of the other components unaltered. By comparing the total number of electrons in two (or several) different measurements, the total excess of ions in the solution could be deduced. In addition to the number of excess ion, this method could also provide a qualitative picture of the ion atmosphere.

1.3 Motivation and Organization of the dissertation

As discussed above, water and ion distributions around biomolecules are very important for their stability and function. A detailed description at the atomic level of hydration structure and ion atmosphere has long been sought. This dissertation devotes to characterize the water and ion distribution around proteins and nucleic acids by combining computational methods and X-ray scattering experiment. The knowledge gained could be used to benchmark current theoretical models, eventually to make improvements in how those theories treating water molecules, ions and cosolvents in general.

Chapter 2 gives a brief introduction of the Integral Equation Theory–Reference Interaction Site Model (RISM). Key concepts and equations of the theory are introduced. Some applications of the theory are then discussed, focusing on the hydration free energy calculation and the solvent distribution problem.

Chapter 3 describes the element X-ray scattering theory and a protocol to compute X-ray scattering signals from atomic models of macromolecules. The solvent distribution computed from 3D-RISM alongside solute models is then used to test the protocol. The intensity profiles (which involve no adjustable parameters) match experiment and molecular dynamics simulations up to wide angle for relatively rigid biomolecules. Calculation of nucleic acid intensity profiles using the conformational ensemble obtained from MD simulation in the solution is in better agreement with experiment than those using a single diffraction structure.

Chapter 4 presents a novel method to extract water and ion distribution from X-ray scattering experiment. The method is shown approximate in nature and only applicable to rigid biomolecules. Nevertheless, it is able to extract aspects of water and ion distributions (beyond the total excess numbers) by combining experimental data for the complete system with calculations for the solutes. The correlation between solute-ion or solute-water can be displayed in both Fourier space (as partial intensities) or real space (as interatomic distribution functions). The resulting ion and water distributions are then used to test predictions from 3D-RISM and MD simulation for proteins and a DNA duplex.

Chapter 5 discusses the extension of X-ray scattering to extract partial molar volume and numbers of excess water molecules and ions from experimental data. We explore a possible use of partial molar volume and the number of excess hydration water to recalibrate solute-solvent interaction. A proof-of-concept example is illustrated as we reoptimize the phosphate groups in a DNA duplex and show that the partial molar volume and number of hydration water approach closer to the experiment.

Chapter 6 gives general conclusions and suggests future directions.

Chapter 2

Brief introduction of Integral Equation Theory and Reference Interaction Site Model

2.1 Introduction

Theory of simple liquids is a subject with long history in which important concepts are born (such as pair correlation function, virial coefficient ...) and are successfully applied for hard spheres and monatomic liquids (for a survey see [107]). Given a pairwise potential, the theory allows to calculate $g(r)$, the structure factor $S(q)$ and other thermodynamic properties of the liquid. RISM theory is an extension proposed by Chandler and Andersen to treat molecular liquids. [108] The theory considers a molecule as a set of separate interaction sites constrained by a strong intramolecular correlations to represent chemical bonds. The partial charge distribution in a molecule is later added by the extended RISM theory (or XRISM). [109] However, the dielectric constant predicted by XRISM is too small compared to the experiment, leading to the enforced input of the dielectric constant into the theory (the so-called dielectrically consistent RISM, or DRISM). [110] As both XRISM and DRISM orientationally average all interactions, they are usually referred as 1D-RISM.

When treating biomolecules that are generally complex and have an arbitrary shape, orientational averaging of interaction shows severe limits. This motivates the development

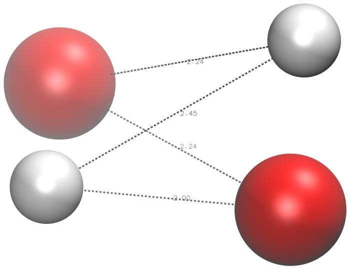


Figure 2.1: Illustration of the interaction-site model for two diatomic “molecules”.

of the so-called 3D-RISM. [111–115] The theory only performs orientational averaging on the solvent, while still treats the solute at a full atomistic 3-dimensional level. The method naturally yields more accurate results compared with pure 1D-RISM and finds a wide application range (from simple ions, electrolyte solutions, ionic liquids to complex proteins and nucleic acids). A wealth of information could be obtained from the theory, including the radial distribution functions, potential of mean forces, solvation free energies, partial molar volumes, compressibilities, etc. (See [116,117]). It should be noted that, however, the solute degree of freedom is ignored by the theory and one must seek a different way to incorporate the solute flexibility, for instance via MD simulation. [118, 119]

In this chapter, we give a very brief introduction of key features of RISM theory. We then discuss the application of the theory, focusing on hydration free energy and solvent distribution around biomolecules.

2.2 Theory

2.2.1 Ornstein–Zernike equation

The integral equation theory starts with the Ornstein–Zernike (OZ) equation, which can be derived from classical theory of liquids (see [41]):

$$h\left(\mathbf{r}_{12},\Omega_1,\Omega_2\right)=c\left(\mathbf{r}_{12},\Omega_1,\Omega_2\right)+\rho\int d\mathbf{r_3}d\Omega_3 c\left(\mathbf{r}_{13},\Omega_1,\Omega_3\right)h\left(\mathbf{r}_{32},\Omega_3,\Omega_2\right) \tag{2.1}$$

where $\mathbf{r}_{ij}$ is the vector connecting particles i and j, Ω_i and Ω_j are the orientation of particles i and j, respectively, relative to $\mathbf{r}_{ij}$, c is the direct correlation function, h is the total correlation function which is related directly to the distribution function g as:

$$h_{ij}\left(\mathbf{r}_{ij},\Omega_i,\Omega_j\right)\equiv g_{ij}\left(\mathbf{r}_{ij},\Omega_i,\Omega_j\right)-1 \tag{2.2}$$

Eq. 2.1 can be rewritten by recursively eliminating h under the integral as:

$$\begin{aligned}h\left(\mathbf{r}_{12},\Omega_1,\Omega_2\right)=c\left(\mathbf{r}_{12},\Omega_1,\Omega_2\right)+\rho\int d\mathbf{r}_3d\Omega_3 c\left(\mathbf{r}_{13},\Omega_1,\Omega_3\right)c\left(\mathbf{r}_{32},\Omega_3,\Omega_2\right)\\+\rho^2\int\int d\mathbf{r}_3 d\varOmega_3 d\mathbf{r}_4 d\varOmega_4 c\left(\mathbf{r}_{13},\varOmega_1,\varOmega_3\right)c\left(\mathbf{r}_{34},\varOmega_3,\varOmega_4\right)c\left(\mathbf{r}_{42},\varOmega_4,\varOmega_2\right)+...\end{aligned} \tag{2.3}$$

The OZ equation effectively defines the direct correlation function c and can be interpreted as following: the total correlation between particles 1 and 2 can be considered as the sum of the direct correlation between those two particles and other indirect parts due to the effect of other particles present in the system. If the liquid is uniform and isotropic, the OZ equation becomes:

$$h\left(r\right)=c\left(r\right)+\rho\int c\left(\left|\mathbf{r}-\mathbf{r}'\right|\right)h\left(r'\right)d\mathbf{r}' \tag{2.4}$$

Taking the Fourier transform of both sides gives the relationship between h and c:

$$\hat{h}\left(k\right)=\frac{\hat{c}\left(k\right)}{1-\rho\hat{c}\left(k\right)} \tag{2.5}$$

2.2.2 Closure equations

In order to solve Eq. 2.1, it is necessary to have a second, so-called closure, equation that relates h and c, which is conventionally written as:

$$g\left(\mathbf{r}_{12},\Omega_1,\Omega_2\right)=\exp\left[-\beta u\left(\mathbf{r}_{12},\Omega_1,\Omega_2\right)+h\left(\mathbf{r}_{12},\Omega_1,\Omega_2\right)-c\left(\mathbf{r}_{12},\Omega_1,\Omega_2\right)+b\left(\mathbf{r}_{12},\Omega_1,\Omega_2\right)\right] \tag{2.6}$$

or in a shorter form

$$g=\exp\left[-\beta u+h-c+b\right] \tag{2.7}$$

Here u is the pair-wise potential energy function and b is an unknown “bridge function”. Finding a good approximation for the bridge function is one important direction of integral equation theory. In the hypernetted-chain approximation (HNC), b is simply set to zero, giving:

$$g_{HNC}=\exp\left[-\beta u+h-c\right] \tag{2.8}$$

The HNC closure gives good results for ionic and polar systems, but poorer results for neutral systems, and it can be difficult to find converged solutions. [116,120,121] To address the convergence issue, Kovalenko and Hirata introduced the KH closure as follows: [122]

$$g_{KH}=\begin{cases}\exp\left[-\beta u+h-c\right] & \text{if } g\leq 1\\ 1-\beta u+h-c & \text{if } g>1\end{cases} \tag{2.9}$$

The partial series expansion of order-n (PSE-n) offers a way to interpolate between KH and HNC, and thus improves the results of KH closure while circumventing the convergence difficulty met in HNC closure: [123]

$$g_{PSE-n} = \begin{cases} \exp\left[-\beta u + h - c\right] & \text{if } g \leq 1 \\ \sum_{i=0}^{n} \frac{\left[-\beta u + h - c\right]^{i}}{i!} & \text{if } g > 1 \end{cases} \tag{2.10}$$

It is obvious that KH is the special case of PSE closure when n=1; and when $n \to \infty$ HNC closure is obtained.

2.2.3 1D-RISM

To apply Eq. 2.1 for use with molecular liquids, the molecule is separated into interaction sites. One critical approximation of RISM theory is to treat the direct correlation function between two molecules $c\left(1,2\right)$ as site-site decomposable:

$$c\left(1,2\right) = \sum_{\alpha\gamma} c_{\alpha\gamma}\left(\left|\mathbf{r}_{\alpha} - \mathbf{r}_{\gamma}\right|\right) \tag{2.11}$$

where α and γ are interaction sites in molecules 1 and 2, respectively. The intramolecular correlation function ω is used to describe the structure of the molecule. For two sites α and β of the same molecule:

$$\omega_{\alpha\beta}\left(r\right) = \frac{\delta\left(r - r_{\alpha\beta}\right)}{4\pi r_{\alpha\beta}^{2}} \tag{2.12}$$

where δ is the Dirac delta function. The geometry of a molecule can be defined if all $r_{\alpha\beta}$ are known. Eq. 2.1 then can be rewritten as: [108, 116, 124]

$$h_{\alpha\gamma}\left(r\right) = \sum_{\lambda}^{N_{site}} \sum_{\beta}^{N_{site}} \omega_{\alpha\lambda}\left(r\right) * c_{\lambda\beta}\left(r\right) * \omega_{\beta\gamma}\left(r\right) + \sum_{\lambda}^{N_{site}} \sum_{\beta}^{N_{site}} \omega_{\alpha\lambda}\left(r\right) * c_{\lambda\beta}\left(r\right) * \rho_{\beta} h_{\beta\gamma}\left(r\right) \tag{2.13}$$

with * is the convolution operator. The equation is usually written in the simpler matrix form

$$\begin{aligned} \boldsymbol{\rho}\mathbf{h}\boldsymbol{\rho} &= \boldsymbol{\omega} * \mathbf{c} * \boldsymbol{\omega} + \boldsymbol{\omega} * \mathbf{c} * \boldsymbol{\rho}\mathbf{h}\boldsymbol{\rho} \\ &= \left[\mathbf{I} - \boldsymbol{\omega} * \mathbf{c}\right]^{-1} \boldsymbol{\omega} * \mathbf{c} * \boldsymbol{\omega} \end{aligned} \tag{\textbf{2.14}}$$

where $\boldsymbol{I}$ is the unit matrix, $\boldsymbol{\rho}$ is a diagonal matrix of density values and $\boldsymbol{\omega}$, $\boldsymbol{h}$, $\boldsymbol{c}$ are matrices of site-site correlation. Eq. 2.13 is then coupled with a set of closure equations to numerically solve for $h_{\alpha\gamma}(r)$. For example, the HNC closure equation for site-site interaction is as following:

$$h_{\alpha\gamma}(r) = \exp\left[-\beta u_{\alpha\gamma}(r) + h_{\alpha\gamma}(r) - c_{\alpha\gamma}(r)\right] - 1 \tag{2.15}$$

where $u_{\alpha\gamma}(r) = \frac{q_\alpha q_\gamma}{r_{\alpha\gamma}} + \varepsilon_{\alpha\gamma}\left[\left(\frac{R_{min,\alpha\gamma}}{r_{\alpha\gamma}}\right)^{12} - 2\left(\frac{R_{min,\alpha\gamma}}{r_{\alpha\gamma}}\right)^{6}\right]$ is the pair interaction potential between two sites, with parameters taken from a molecular mechanics force field.

The dielectric constant computed from the RISM equations is significantly low compared with experiment (for SPC/E water $\epsilon < 20$), causing problems when ions at finite concentration are present in the solvent. The error was presumed to be caused by the long range asymptotics of HNC-like closures. Perkyns and Pettitt introduced a bridge-like correction $\boldsymbol{\zeta}$ into Eq. 2.14: [110,120]

$$\boldsymbol{\rho}\mathbf{h}\boldsymbol{\rho} = \left[\mathbf{I} - \boldsymbol{\omega'} * \mathbf{c}\right]^{-1} \boldsymbol{\omega'} * \mathbf{c} * \boldsymbol{\omega'} + \boldsymbol{\zeta} \tag{2.16}$$

with $\boldsymbol{\omega'} = \boldsymbol{\omega} + \boldsymbol{\zeta}$. The $\boldsymbol{\zeta}$ matrix is determined by the input dielectric constant.

2.2.4 3D-RISM

The 3D-RISM considers a solution consisting of a molecular solvent and a single solute. Similarly for 1D-RISM, one also has to assume that the solute-solvent direct correlation function $c^{uv}(r_{12}, \Omega_1, \Omega_2)$ can be decomposed into contributions of sites α of solvent molecule 2:

$$c^{uv}(r_{12}, \Omega_1, \Omega_2) = \sum_\alpha c_\alpha^{uv}(\mathbf{r}_{1\alpha}) \tag{2.17}$$

For a vanishing solute density, the RISM-OZ equation splits up into the equation for pure solvent and solute-solvent (with superscripts u and v denote solute and solvent, respectively):

$$h_{\alpha\gamma}^{vv}(\mathbf{r}_\alpha, \mathbf{r}_\gamma) = c_{\alpha\gamma}^{vv}(\mathbf{r}_\alpha, \mathbf{r}_\gamma) + \sum_\lambda \rho_\lambda^v \int c_{\alpha\lambda}^{vv}(\mathbf{r}_\alpha, \mathbf{r}_\lambda)\, h_{\lambda\gamma}^{vv}(\mathbf{r}_\lambda, \mathbf{r}_\gamma)\, d\mathbf{r}_\lambda \tag{2.18}$$

$$h_\alpha^{uv}(\mathbf{r}_s, \mathbf{r}_\alpha) = c_\alpha^{uv}(\mathbf{r}_s, \mathbf{r}_\alpha) + \sum_\lambda \rho_\lambda^v \int c_\lambda^{uv}(\mathbf{r}_s, \mathbf{r}_\lambda)\, h_{\lambda\alpha}^{vv}(\mathbf{r}_\lambda, \mathbf{r}_\alpha)\, d\mathbf{r}_\lambda \tag{2.19}$$

Eq. 2.18 can be well approximated by Eq. 2.16. One can couple it with a set of closure equations to obtain the solvent susceptibility $\chi_{\alpha\gamma}^{vv}(r)$, which contains all information about the bulk solvent:

$$\chi_{\alpha\gamma}^{vv}\left(r\right)=\omega_{\alpha\gamma}^{vv}\left(r\right)+\rho_{\alpha}^{v}h_{\alpha\gamma}^{vv}\left(r\right)\tag{2.20}$$

The solvent susceptibility will be used (coupled with a set of 3D closure equations) to numerically solve the so-called 3D-RISM equation:

$$h_{\alpha}^{uv}\left(\mathbf{r}\right)=\sum_{\lambda}\int c_{\lambda}^{uv}\left(\mathbf{r}-\mathbf{r}'\right)\chi_{\lambda\alpha}^{uv}\left(r'\right)d\mathbf{r}'\tag{2.21}$$

For example, the HNC closure for 3D is as following:

$$h_{\alpha}^{uv}\left(\mathbf{r}\right)=\exp\left[-\beta u_{\alpha}^{uv}\left(\mathbf{r}\right)+h_{\alpha}^{uv}\left(\mathbf{r}\right)-c_{\alpha}^{uv}\left(\mathbf{r}\right)\right]-1\tag{2.22}$$

with $u_{\alpha}^{uv}\left(\mathbf{r}\right)$ is the interaction potential between the solute molecule and the α site of the solvent, computed as a superposition of the site-site interaction between solute sites and α site $u_{\alpha}^{uv}\left(\mathbf{r}\right)=\sum_{a}^{M}u_{a\alpha}\left(\left|\mathbf{r}_{a}-\mathbf{r}\right|\right)$.

2.3 Application of RISM theory

Here we only focus on the calculation of solvation free energy, partial molar volume of the solute and the distribution of solvents around the solute. There are other works exist in the literature including hybridization of RISM with MD simulation, quantum mechanics or Monte Carlo (see a recent review by Fedorov and coworkers [117]).

2.3.1 Solvation free energy

2.3.1.1 Analytical expression

One of the most important quantities can be computed directly from 3D-RISM is the solvation free energy of the solute (or the excess chemical potential in infinite dilution). In RISM, the Kirkwood formula can be used to calculate the excess chemical potential from the radial distribution function:

$$\Delta G_{solv}=\int_{0}^{1}d\lambda\left\langle\frac{\partial U\left(\mathbf{r}_{1},\mathbf{r}_{2},...,\mathbf{r}_{N},\lambda\right)}{\partial\lambda}\right\rangle_{\lambda}\tag{2.23}$$

which can be rewritten as for the solute site α: [125]

$$\Delta\mu_{\alpha}=\sum_{\gamma}\rho_{\gamma}\int_{0}^{1}\left[\int\frac{\partial u_{\alpha\gamma}\left(r;\lambda\right)}{\partial\lambda}g_{\alpha\gamma}\left(\mathbf{r};\lambda\right)d\mathbf{r}\right]d\lambda\tag{2.24}$$

For HNC–like closures, the above equation can be solved analytically: [122,123]

$$\Delta\mu_{\alpha} = k_B T \sum_{\gamma} \rho_{\gamma} \int \left[\frac{h_{\alpha\gamma}^2}{2} - c_{\alpha\gamma} - \frac{h_{\alpha\gamma} c_{\alpha\gamma}}{2} - \frac{\left(t_{\alpha\gamma}^*\right)^{n+1}}{(n+1)\,!} \Theta\left(t_{\alpha\gamma}^*\right) \right] d\mathbf{r} \tag{2.25}$$

where $t_{\alpha\gamma}^* = -\beta u_{\alpha\gamma} + h_{\alpha\gamma} - c_{\alpha\gamma}$ and Θ is the Heaviside step function ($\Theta(x) = 1$ for $x > 0$ and $\Theta(x) = 0$ otherwise). For $n \to \infty$ the last term of the integrand vanishes and one obtains the result for HNC closure. For $n = 1$, the equation gives KH $\Delta\mu$.

2.3.1.2 Correction for $\Delta\mu$

It is known that the solvation free energy calculated by 3D-RISM is not in agreement with the experiment. Palmer *et al.* recognize that the error of the predicted $\Delta\mu$ correlates very well with the calculated partial molar volume $\overline{V}$ by 3D-RISM and propose an empirical formula to adjust the output free energy: [126]

$$\Delta\mu^{UC} = \Delta\mu + a_1 \rho \overline{V} + a_0 \tag{2.26}$$

where a_0 and a_1 are two parameters obtained by pre-fitting small molecule hydration free energies to experiment and ρ is the solvent bulk density.

Later, it is shown that 3D-RISM predicts the electrostatic component of the HFE accurately but requires a modification of the non-polar contribution. [127] By scaling only the direct correlation function of solvent inside the excluded volume of the solute, the authors propose the so-called cavity-corrected functional to compute the HFE from 3D-RISM:

$$\Delta\mu^{CC} = \Delta\mu + \frac{1}{2} k_B T \rho \left(1 - \gamma\right) \int_{exclV} c_o^{np}\left(\mathbf{r}\right) d\mathbf{r} \tag{2.27}$$

Here, $c_o^{np}(\mathbf{r})$ is the non-polar direct correlation of oxygen water which is calculated by setting the solute charge to zero. The integral is only over the excluded volume of the solute. This strategy leaves a single parameter remained, γ, which can be fitted by MD simulations of simple Lennard–Jones solutes.

In a different approach, Borgis and colleagues offer an explanation to the partial molar volume dependence of the calculated HFE errors. [128] This somewhat relates to the appreciated thermodynamic inconsistency of approximate theories such as RISM. Basically, the pressure in the system can be calculated by three different routes: via the virial equation, via the compressibility equation and via the Helmholtz free energy (see Appendix 6). [41] Values computed from different routes are often widely different from each other, and thus the theory is considered thermodynamically inconsistent. The pressure computed for water from RISM is around 3–4 orders of magnitude larger than experiment. [121] The high pressure affects the HFE of the solute since:

$$\Delta\mu = \Delta U - T\Delta S + P\Delta V \tag{2.28}$$

where U, T, S, P, V are the internal energy, temperature, entropy, pressure and volume, respectively. The authors thus propose an *ad hoc* protocol to replace the incorrect pressure by the experimental value to correct the HFE:

$$\Delta\mu^{P} = \Delta\mu - \rho k_B T \left[2 - \frac{\rho}{2}\hat{c}_s\left(0\right)\right]\Delta V \tag{2.29}$$

with $\hat{c}_s\left(0\right) = \int c_s\left(|\mathbf{r}|\right)d\mathbf{r}$, c_s is the solvent direct correlation function and ΔV is the partial molar volume of the solute. At a first glance, it seems that Eq. 2.27 is similar to Eq. 2.29 and the two methods should be related. However, more careful inspection shows that $c_o^{np}\left(\mathbf{r}\right)$ in Eq. 2.27 is the solute-solvent correlation while $c_s\left(r\right)$ in Eq. 2.29 is the solvent-solvent correlation. It is currently not clear whether the two approaches have any deeper relation. Nonetheless, given that the partial molar volume directly relates to the solute-solvent total correlation function $c\left(\mathbf{r}\right)$ (see Section 5.3, Eq. 5.21), it is therefore expected there should be a link between the two correction methods. Eq. 2.29 could be recast in terms of the solvent isothermal compressibility χ_T as:

$$\Delta\mu^{P} = \Delta\mu - \frac{1}{2}\rho k_B T\left(3\rho k_B T\chi_T + 1\right)\Delta V \tag{2.30}$$

Borgis *et al.* also suggest that the pressure computed by RISM should take the following form:

$$\begin{aligned} P &= \frac{n_s+1}{2}\rho k_B T - \frac{1}{2}k_B T\rho^2\hat{c}\left(0\right) \\ &= \frac{n_s}{2}\rho k_B T + \frac{1}{2\chi_T} \end{aligned} \tag{2.31}$$

where n_s is the number of site in the solvent molecule ($n_s = 3$ for water).

2.3.2 Ion and water distribution

3D-RISM allows the equilibrium density distribution of solvents (including water, ions and cosolvents in general) around an arbitrary solute to be computed at a significantly reduced computational cost compared to explicit simulation. Since the 3-dimensional distribution of solvent is the direct output from the theory, it should be straightforward to compare the theory performance with experiment and explicit simulation.

Early works from Hirata and coworkers illustrate that 3D-RISM is capable of detecting both water and ion binding sites in proteins with a reasonably good accuracy. [8, 129] The method is later expanded to small molecules and ions (such as H_3O^+, Ne, NO, CO_2, NH_3, urea, glycerol ...) to study their transport characterization through channels by computing their spatial distribution and mapping the potentials of mean force. [130, 131] In principle, any solvent, regardless how complex they are, should work with 3D-RISM. However, when the number of interaction sites in the solvent increases (due to cosolvents and/or the complexity of the solvent), the solution of RISM equations gets harder to find. It is currently not clear where the problem is. An interesting way to deal with a complex solvent is to perform separate calculations for fragments of the solvent (*i.e.* treating the fragments as unique species), and then combine the spatial distributions in a way that maintains the molecular solvent structure. [132] This approach shows much promise in *in silico* docking, removing the need of an empirical scoring function that is still widely used today.

There are still questions that remained: how well does 3D-RISM hydration structure and ion cloud compared to the much more expensive explicit MD simulation and ultimately, to experimental data? Stumpe *et al.* perform a benchmark calculation to compare 3D-RISM water structure with those from MD. [133] It has been shown that the water distribution functions from two methods are very similar for a model protein. In a different work, 3D-RISM is also shown to accurately predict hydration patterns of a nucleic acid. [134] The spine of hydration in the minor grooves is found to be similar with MD simulation. 3D-RISM also has the ability to discriminate different bases and their hydration structure. In a recent study, the ion atmosphere around nucleic acids from 3D-RISM is predicted as

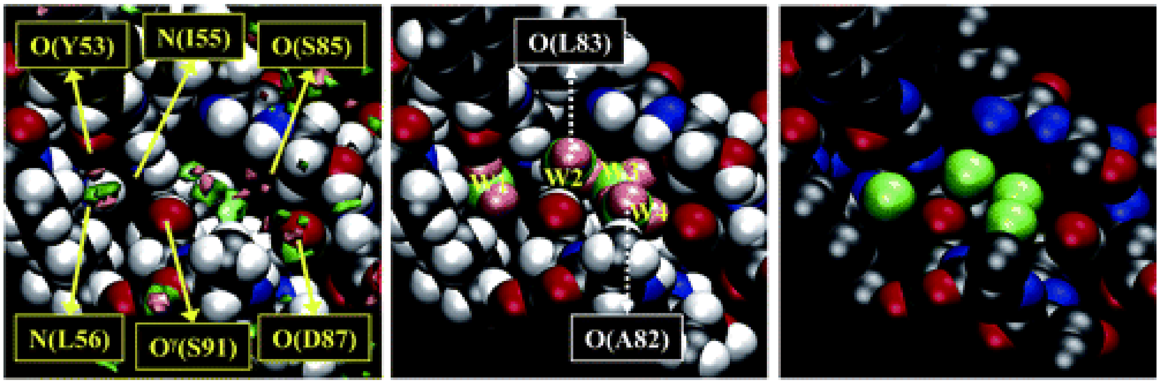


Figure 2.2: Water distribution in the cavity of hen egg-white lysozyme calculated with 3D-RISM compared with the crystallographic water sites (right). Left: the isosurface of water oxygen (green) and hydrogen (pink) distribution functions with isovalue > 8. Center: the most probable model of the hydration structure reconstructed from the isosurface plots. Reprinted with permission from Imai *et al.* [8] Copyright 2005 American Chemical Society.

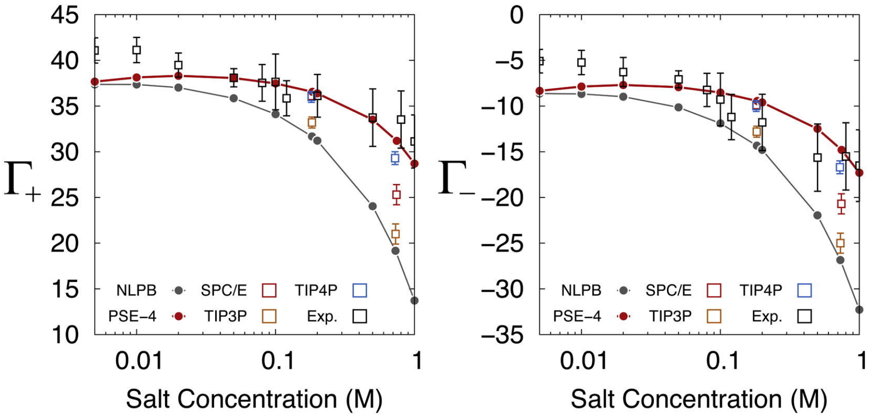


Figure 2.3: Comparison between theoretically predicted numbers of excess Na$^+$ (left) and Cl$^-$ (right) and experimental ion counting results. Reprinted from Giambasu *et al.* [9], with permission from Elsevier.

multilayer structure and in overall the shape is in very good agreement with MD simulation. Quantitative comparison of the number of excess ions around the DNA computed by the theory is significantly better than those from continuum models. Similarly to MD simulation, 3D-RISM performs best around physiological concentration of salts and slightly underestimates the counterion accumulation at high concentrations. The ability to model very diluted solutions is one of the great asset of the theory, since MD simulation for those is extremely costly to perform.

Chapter 3

Calculation of X-ray scattering profiles from atomic models

3.1 Introduction

Small angle X-ray scattering (SAXS) is a widely used diffraction tool to study biomolecule structure. It shares very much the same governing principles with another closely related method X-ray crystallography. While X-ray crystallography is used for structure determination of crystals, SAXS probes the molecular structure in the solution. X-ray scattering signals are known to depend greatly on the protein-solvent interface and interaction. The computation of scattering profiles from atomic models can be a difficult task, even in the simplest case where the solute molecule adopts a known, single and relatively rigid conformation in solution. Because both solute and solvent contribute to the scattering, the perturbation of the solvent (usually water and ions) by the biomolecule must be understood and properly modeled in order to make comparisons to experiment.

Several methods have been developed to include the contribution of water to the overall scattering profiles. Most rely on the simplified models of water to account for the scattering of excluded volume and hydration shell. CRYSOL, for example, assumes a layer of uniform excess hydration density around the surface of the protein. [135] However, the surface topology, electrostatics and hydrophobicity patterns surely play a role as well. The water shell, additionally, is composed of successive layers of excess and deficient density relative to the bulk. Different approaches have been considered to describe the hydration shell more realistically, from treating solvent as an assembly of freely orienting and interacting dipoles, [136] to reconstructing the three-dimensional hydration shell by combining a set of proximal radial distribution functions for different atom types extracted from MD simulations. [137] In principle, MD simulations can also provide such information, but these are difficult to converge (especially for ions in the vicinity of charged biomolecules), and are computationally tedious and expensive.

The promising intermediate approach explored here uses integral equation theory to estimate thermally averaged water and ion distributions on a 3-dimensional grid surrounding the biomolecule at a fraction of the cost of MD simulations. These estimates are certainly imperfect ones, as they are based on simple force field models for the relevant atomic interactions, and use approximate closures and averaging procedures to treat molecular solvents like water. We show here that they are nevertheless accurate enough to provide good estimates of X-ray scattering out to angles corresponding to $q < 1.5$ Å^{-1} ($q = 4\pi \sin\theta/\lambda$ where 2θ is the scattering angle and λ is the X-ray wavelength). Once the force field and closure are determined, there are no adjustable parameters in this model. The results can be of particular use in treating salt solutions and for comparisons to experiments where an absolute calibration near $q = 0$ is available. We do not consider here the "inverse" problem of interpreting experimental data arising from an unknown structure, or from samples where an ensemble of structures is contributing to the scattering. Nevertheless, the computations described here are efficient enough to be applied to large numbers of proposed structures (or to structural ensembles), and are based on a physically-motivated model for solvent effects that appears to be more accurate than any other currently-available procedure. It is likely that these ideas could form the basis for model discovery and selection in a wide range of problems.

The chapter first starts with some basic theory of diffraction and scattering of X-ray as well as a very brief description of X-ray experiment. We then give an overview of different methods in the literature to calculate X-ray scattering profiles from atomic models. Next, we present our method based on the integral equation theory and make comparisons to both experiment and MD simulations for relatively rigid proteins, several nucleic acid duplexes and a dozen of biomolecules taken from the *BioIsis.net* database. The intensity profiles calculated from our method (which involve no adjustable parameters) match experiment and MD simulations up to wide angle. We illustrate the importance of using an ensemble structure for calculation of nucleic acid profiles rather than using a single fiber-diffraction model. In cases where an absolute calibration of the experimental data at $q = 0$ is available, we show that numbers of excess water and ions can be extracted from the experimental data.

3.2 Principle of X-ray scattering

3.2.1 Scattering of X-ray by an electron

X-rays are high energy electromagnetic beam and therefore the interaction of X-ray with matter is dominated almost completely by the interaction with electrons. There are two different processes involved in the scattering of the incident radiation by a free electron:

1. Elastic (or coherent or Thomson) scattering: the photon does not lose energy and the emitted photon has the same frequency as the incoming photon.

2. Inelastic (or incoherent or Compton) scattering: the electron accepts some momentum from the incoming photon and the emitted photon has smaller frequency (or longer wavelength) compared with the incoming photon.

The latter is very weak at small angles and is neglected in SAXS. The scattered intensity I_e is therefore can be computed from the incoming intensity I_0 by Thomson formula:

$$I_e\left(\theta\right) = I_0 \frac{e^4}{m^2 c^4 r^2}\left(\frac{1+\cos^2 2\theta}{2}\right) \tag{3.1}$$

with e, m are the charge and mass of the electron, respectively, and r is the distance from the electron to the point of observation. The term in the bracket is practically equal to 1 for the small angles used in SAXS, and thus I_e is considered a constant. For the sake of brevity, this intensity and the magnitude of its amplitude will be set to 1 from now on.

3.2.2 Scattering of X-ray by an atom

As the energy of an X-ray photon is very large compared to the binding energy of an atom (except heavy atom, which relates to the so-called anomalous scattering), all electrons are effectively free (*i.e.* not influenced by the nucleus). Every electron becomes the source of a scattered waves with the same intensity but with different phases. Thus the atomic scattering factor can be computed by summing up all contributions from the electron cloud:

$$f\left(\mathbf{q}\right) = \int \rho\left(\mathbf{r}\right) \exp\left(-i\mathbf{q}.\mathbf{r}\right) d\mathbf{r} \tag{3.2}$$

with $\rho\left(\mathbf{r}\right)$ is the electron density, and $\boldsymbol{q}$ is the momentum transfer with its magnitude relates to the X-ray wavelength λ and the scattering angle 2θ as following $q = \frac{4\pi}{\lambda}\sin\theta$. The atomic scattering factor is thus the Fourier transform of the electron density. In the case of forward scattering ($q = 0$), then $f\left(0\right) = Z$. Eq. 3.2 can also be written as following, considering the electron density is spherically symmetric (using $\langle \exp\left(-i\mathbf{q}.\mathbf{r}\right)\rangle = \frac{\sin(qr)}{qr}$):

$$f\left(q\right) = 4\pi \int \rho\left(r\right) \frac{\sin\left(qr\right)}{\left(qr\right)} r^2 dr \tag{3.3}$$

The atomic scattering factor can be computed from first principles from electronic wave functions. In practice, it is desirable to have an empirical formula for fast evaluation of the atomic factor. Cromer and Mann fit $f\left(q\right)$ as the sum of 4 Gaussian functions and a constant c: [138]

$$f\left(q\right) = \sum_{i}^{4} a_i \exp\left(-b_i \frac{q^2}{16\pi^2}\right) + c \tag{3.4}$$

with a and b are the tabulated constants (available for atoms and monatomic ions). Others suggest to get rid of the constant c, and instead fit $f\left(q\right)$ with N Gaussians (with N determined based on the desired accuracy). [139]

3.2.3 Scattering of X-ray by a molecule

We now proceed to the calculation of X-ray scattering for a molecule in vacuum. Similarly to the atomic scattering factor, the amplitude of the scattering wave from a molecule can be computed as a superposition of all partial waves from all atoms in the molecule:

$$F\left(\mathbf{q}\right) = \sum_{\alpha} f_\alpha\left(q\right) \exp\left(-i\mathbf{q}.\mathbf{r}_\alpha\right) \tag{3.5}$$

where $f_\alpha\left(q\right)$ is the atomic scattering factor of atom α. For randomly oriented molecules, the scattering intensity needs to be averaged all over every direction:

$$\begin{aligned} I\left(q\right) &= \frac{1}{4\pi}\int_{\Omega} F\left(\mathbf{q}\right)F^{*}\left(\mathbf{q}\right)d\Omega \\ &= \sum_{\alpha}\sum_{\beta} f_{\alpha}\left(q\right)f_{\beta}\left(q\right)\frac{\sin\left(qr_{\alpha\beta}\right)}{qr_{\alpha\beta}} \end{aligned} \tag{3.6}$$

where $F^{*}\left(\mathbf{q}\right)$ is the complex conjugate of $F\left(\mathbf{q}\right)$. Eq. 3.6 is the well-known Debye's formula to calculate the scattering intensity.

3.2.4 Anomalous scattering

If the X-ray beam energy is very close to the binding energy of the atom then the situation becomes more complex. Some photons are scattered normally. Some photons are absorbed and re-emitted at lower energy (fluorescence). Some photons are absorbed and re-emitted at the same energy, however with a gain to its phase (*i.e.* it is retarded compared to a normally scattered photon). The atomic scattering factor becomes a complex number:

$$f\left(q,E\right) = f_{0}\left(q\right) + f'\left(E\right) + f''\left(E\right)i \tag{3.7}$$

where $f_{0}\left(q\right)$ is the regular atomic scattering factor discussed above; $f'\left(E\right)$ and $f''\left(E\right)$ are the real and imaginary parts, respectively, of the anomalous scattering factor. They are practically independent of the scattering angle (or q) and only depend on the X-ray energy E. The imaginary part is proportional to the atomic absorption coefficient μ and therefore could be determined experimentally:

$$f''\left(E\right) = \frac{\pi mc}{2e^{2}h}E\mu\left(E\right) \tag{3.8}$$

The real part can be obtained by the Kramers–Kronig equation (or Hilbert transform):

$$f'\left(E\right) = \frac{2}{\pi}\int_{0}^{\infty}\frac{E'f''\left(E'\right)}{E^{2}-E'^{2}}dE' \tag{3.9}$$

As one uses lower and lower X-ray energy beam, the X-ray absorption of water increases and therefore preventing the use of anomalous scattering for biologically relevant elements such as O, N, C, Na, Mg, K. Figure 3.1 shows the anomalous scattering factors of Rb and Sr as an example. Note that f' is always negative and f'' is very close to zero near the absorption edge. Therefore, using the X-ray beam with the energy right below the element absorption edge is equivalent to reducing the "effective" number of electrons in that element. This fact will be used extensively to study the ion distribution and will be discussed later.

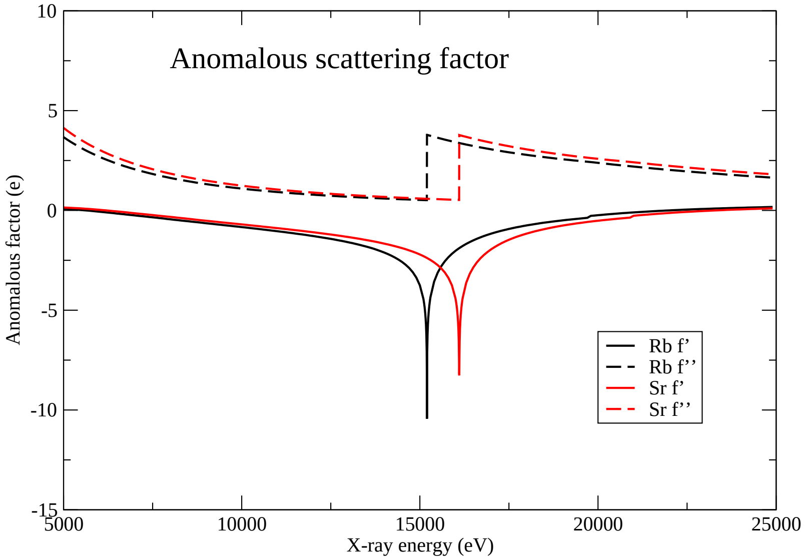
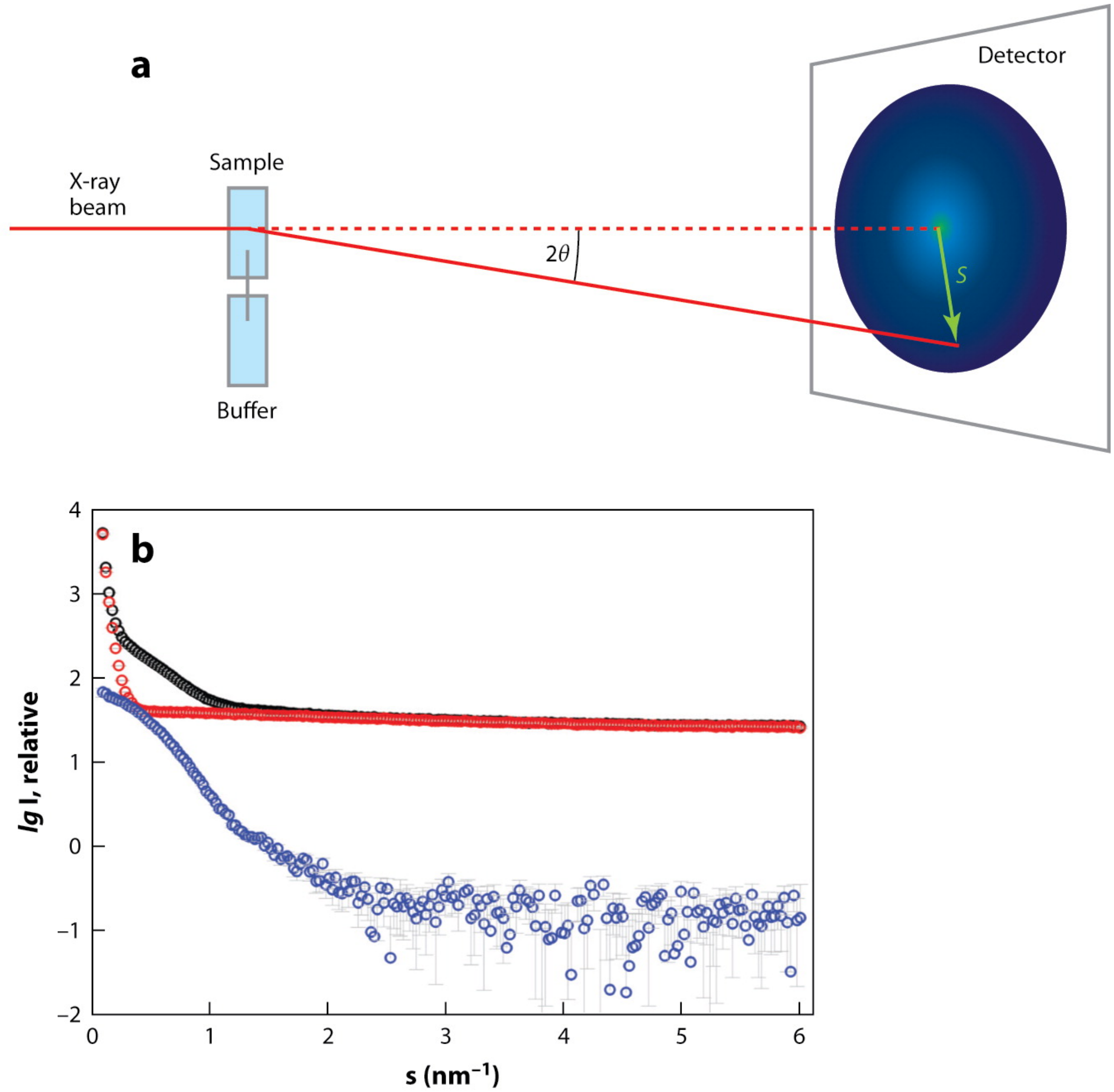


Figure 3.1: Anomalous scattering factors for Rb and Sr at different X-ray beam energy. Data from “http://skuld.bmsc.washington.edu/scatter/AS_form.html” (accessed on Nov 21st, 2015).

3.2.5 Experimental measurement of biomolecule X-ray scattering

Before attempting to calculate the X-ray scattering signal, it is useful to understand what is measured experimentally and how the experiment is carried out. Here we give a very brief sketch of X-ray scattering experiment. For a more in-depth description, we refer to some excellent books and reviews elsewhere. [140–144]

Solution X-ray scattering experiment is conceptually simple as schematically shown in Figure 3.2. A monochromatic X-ray beam is brought to a sample from which only a small amount of the X-ray scatters, while most pass through the sample without interacting with it. The X-ray scattering pattern is then detected at a detector by counting how many photons reach the detector. The 2-dimensional pattern is then radially averaged to obtain the sample scattering curve. A similar experiment is repeated with the sample replaced by the pure solvent. The solvent scattering curve is subsequently subtracted from the sample scattering profile to obtain the scattering profile caused by only the macromolecules. The forward scattering signal ($q = 0$) cannot be detected in the experiment due to the strong



Blanchet CE, Svergun DI. 2013.
Annu. Rev. Phys. Chem. 64:37–54

Figure 3.2: Scheme of SAXS experiment. a) An X-ray beam targets the sample and the scattered photons are collected on a detector. The signal is radially averaged to get the scattering curve. b) Two separate measurements are carried out: the “sample” with the protein present (black curve) and the “buffer” with only pure solvent and/or buffer (red curve). The scattering pattern is the difference between those two curves (blue curve). Figure taken from Blanchet and Svergun. [10]

non-interacting forward beam, instead it must be obtained by extrapolation.

3.3 Calculation of X-ray scattering signals from atomic models

As in section 3.2.3, we so far only consider the calculation of the molecule in vacuum. Since the experimental scattering profile is the difference of the scattering data between the sample and solvent, one also needs to take into account of the solvent. The solvent affects the solute scattering signal in two ways:

- Reduce the scattering contrast of the solute since the background now is bulk water with the electron density of 0.333 e/Å^3. The average electron density of a globular protein is around 0.44 e/Å^3. The scattering contrast of a protein in water is, therefore, only $^1/_4$ compared with that in gas phase.
- Contribute to the scattering signal through the solvation shell. It is known both experimentally and theoretically that the water density around the protein is higher than the bulk value. [38, 39] The contribution from the solvation shell thus needs to be taken into account to achieve a good accuracy.

It is obvious that the calculation of X-ray scattering for molecules in vacuum is straightforward. The most challenged part is how to incorporate the solvent effect into the computation. We here give a brief review of current widely used methods to calculate X-ray scattering intensity of biomolecules in solution, focusing on their way of treating solvation shell. Next we present our method using 3D-RISM as the hydration model and illustrate that our method is better than current competing models.

3.3.1 Overview of different methods

It is worth noting that most (if not all) methods found in the literature involve adjustable parameters to fit the calculated profiles with experimental data. The omnipresence of parameters highlight the lack of a good hydration model that can be used to compute SAXS signals in the past. Recently, using of explicit MD simulation allows SAXS profiles to be computed at a very good accuracy without fitting parameters. However, MD simulation is quite expensive since two separate simulations need to be performed: one for the "sample" with proteins and another for pure solvent. Here we review different approaches for SAXS calculation, emphasizing the hydration shell modeling. Other different methods exist such as: coarse-graining the solute, using different quadrature and approximate protocols to evaluate the spherical averaging, etc. [145]

3.3.1.1 CRYSOL

CRYSOL is, arguably, the most successful tool for evaluating SAXS profiles for bio-molecules in solution and is still a *de facto* standard today due to its simplicity and fast execution. [135] The hydration shell in CRYSOL is approximated by a border layer with a uniform excess density $\Delta\rho$ relative to the bulk value (see Figure 3.3). The thickness of that hydration layer is kept constant at 3 Å while $\Delta\rho$ is allowed to vary. The intensity is calculated as:

$$I(q) = \langle F(\mathbf{q}) F^{*}(\mathbf{q}) \rangle_{\Omega} \tag{3.10}$$

with the excess scattering amplitude:

$$F(\mathbf{q}) = F_{vac}(\mathbf{q}) - \rho F_{c}(\mathbf{q}) + \Delta\rho F_{hyd}(\mathbf{q}) \tag{3.11}$$

where $F_{vac}(\mathbf{q})$ is the scattering amplitude from the molecule in vacuo, $F_{c}(\mathbf{q})$ and $F_{hyd}(\mathbf{q})$ are the scattering amplitudes from the excluded volume and the hydration layer, respectively. The excluded term is calculated as the form factor of a ghost-solute, where everything is similar to the real solute (radius and position of the atoms) except the scattering contrast is of the bulk solvent. Multipole expansion is used to carry out the spherical averaging in Eq. 3.10, leading to the working equation to compute SAXS profiles:

$$I(q) = \sum_{l=0}^{L} \sum_{m=-l}^{l} \left[A_{lm}(q) - \rho C_{lm}(q) + \Delta\rho B_{lm}(q)\right]^{2} \tag{3.12}$$

where the truncation L determines the accuracy of the method; $A(q)$, $B(q)$, $C(q)$ are the multipole expansion coefficients of $F_{vac}(\mathbf{q})$, $F_{hyd}(\mathbf{q})$ and $F_{c}(\mathbf{q})$, respectively. The intensity can be fitted to experiment using two parameters: (i) the contrast of the hydration shell $\Delta\rho$ and (ii) the effective atomic radius of the biomolecule (for calculating the contribution of the excluded volume $C(q)$). Additionally, the overall scaling factor is also needed if one compares with absolute intensity data.

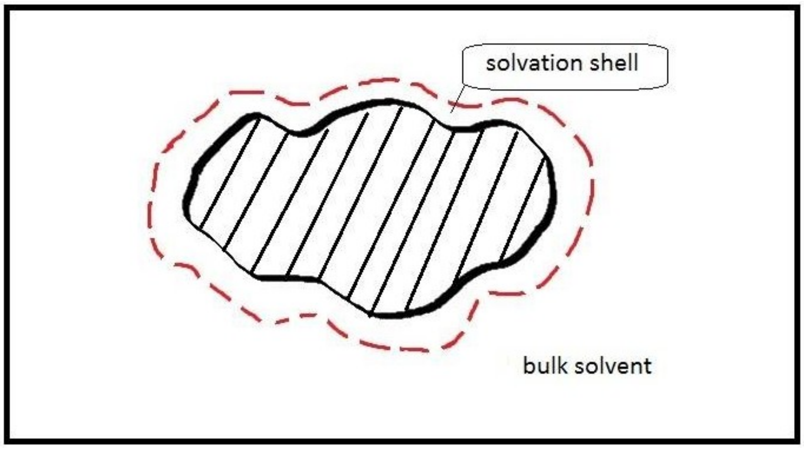


Figure 3.3: Illustration of a protein in aqueous solution and its solvation shell.

3.3.1.2 FoXS

FoXS uses the Debye's formula for computing SAXS profiles: [146, 147]

$$I\left(q\right)=\sum_{i}^{N}\sum_{j}^{N}f_{i}\left(q\right)f_{j}\left(q\right)\frac{\sin\left(qd_{ij}\right)}{qd_{ij}} \tag{3.13}$$

where d_{ij} is the distance between atoms i and j, N is the number of atoms in the molecule. The solvent effect is implicitly incorporated into the model by modifying the atomic scattering factor as:

$$f\left(q\right)=f_{v}\left(q\right)-c_{1}f_{s}\left(q\right)+c_{2}sf_{w}\left(q\right) \tag{3.14}$$

with $f_v\left(q\right)$ is the regular atomic scattering factor *in vacuo*, $f_s\left(q\right)$ is the form factor of the dummy atom that represents the displaced solvent, s the fraction of solvent accessible surface of the atom and $f_w\left(q\right)$ is the water form factor. Aside from the overall scaling constant, two parameters exist in the model: c_1 to adjust the total excluded volume of the atom and c_2 to modify the density of water in the hydration layer.

3.3.1.3 AXES

Instead of treating the hydration shell implicitly, Bax and colleagues suggest to use explicit water molecules to model the hydration shell. [148] A pure water box is pre-equilibrated by MD simulation and N snapshots are extracted after the water density reaches equilibrium. The biomolecule is then placed into those boxes and water molecules that are too far away from the biomolecule are removed (only keep water within 3 Å from the solute). The remained water is then separated into two set: the displaced set including water molecules that clash with the solute and the surface set including the rest of water molecules. The SAXS profile is then computed by averaging the intensity over all orientations and N snapshots of the equilibrated solvent box:

$$I\left(q\right)=\left\langle\left\langle F\left(\mathbf{q}\right)F^{*}\left(\mathbf{q}\right)\right\rangle_{\Omega}\right\rangle_{N} \tag{3.15}$$

$$F\left(\mathbf{q}\right)=F_{vac}\left(\mathbf{q}\right)-F_{disp}\left(\mathbf{q}\right)+\Delta\rho F_{surf}\left(\mathbf{q}\right) \tag{3.16}$$

where $F_{vac}\left(\mathbf{q}\right)$, $F_{disp}\left(\mathbf{q}\right)$ and $F_{surf}\left(\mathbf{q}\right)$ are the form factor of the biomolecule *in vacuo*, the water molecules in the displaced set and surface set, respectively. The model uses two adjustable parameters to fit with experiment: $\Delta\rho$ to scale the scattering contrast of water molecules near the surface of the protein and another one (not discussed here) to take into account the source of experimental data variability. If absolutely calibrated profiles are available, a scaling factor is also needed.

3.3.1.4 AquaSAXS

AquaSAXS represents solvent distribution around a biomolecule using a 3-dimensional grid. [136] The solvent is no longer treated as a homogeneous dielectric medium but modeled as self-orienting interacting dipoles with varying density and dielectric constant, the so-called Poisson–Boltzmann–Langevin formalism. [44, 149] The excess scattering amplitude is computed as:

$$F\left(\mathbf{q}\right)=F_{vac}\left(\mathbf{q}\right)-\rho F_{sev}\left(\mathbf{q}\right)+\rho F_{hyd}\left(\mathbf{q}\right) \tag{3.17}$$

where $F_{sev}\left(\mathbf{q}\right)$ is the form factor of the excluded volume, calculated by the ghost-solute approach, similarly to CRYSOL. The hydration shell contribution is evaluated by summing all the excess density of the solvent as following:

$$F_{hyd}\left(\mathbf{q}\right)=c\sum_{j}^{N_{grid}}a^{3}\left[g\left(\mathbf{r}\right)-1\right]\exp\left(-i\mathbf{q}.\mathbf{r}_{j}\right) \tag{3.18}$$

with $g\left(\mathbf{r}\right)$ is the normalized water density and a is the grid resolution. The AquaSAXS approach still needs to keep three parameters to adjust the agreement with experimental data: the effective atomic radius (as in CRYSOL for computing the excluded term), the constant c in Eq. 3.18 to reflect the contribution of the hydration term and the overall scaling constant.

3.3.1.5 HyPred

Originating from the works of Pettitt and colleagues, [150] Virtanen *et al.* construct a set of proximal radial distribution function (pRDF) for different types of atoms in proteins by explicit simulation. [151] Those pRDFs are then used to generate the solvent distribution around a biomolecule for SAXS calculation. [137] Electrons are "distributed" into a 3-dimensional grid based on those pRDFs and the contribution of the grid is determined by (which is known as the CUBE method [152]):

$$F_{grid}\left(\mathbf{q}\right) = \sum_{n}^{N} 8\Delta\rho \frac{\sin\left(\frac{q_x a}{2}\right)\sin\left(\frac{q_y a}{2}\right)\sin\left(\frac{q_z a}{2}\right)}{q_x q_y q_z} \exp\left(-i\mathbf{q}.\mathbf{r}_n\right) \tag{3.19}$$

where $\Delta\rho$ is the excess electron density at each grid point. The scattering amplitude for the solute and then the total X-ray intensity are evaluated similarly as in other methods. It is not reported in their paper how many fitting parameters are used in the calculation, but we show later that at least the overall scaling constant is needed to compare with absolutely calibrated experimental data.

3.3.2 SAXS profile calculation based on 3D-RISM

3.3.2.1 Formulation

Here, we propose another method to calculate SAXS profiles for biomolecules. The solvent distribution (including water and ions) is computed by employing 3D-RISM. X-ray profiles are then calculated from this distribution alongside with the solute geometry.

We first summarize the derivation from Park *et al.* to compute X-ray scattering curve from MD simulations. [11] Other relevant works include [153–156]. The electron density of the system $\tilde{A}\left(\mathbf{r}\right)$ is separated into contribution from the solute plus its hydration shells $\tilde{A}_1\left(\mathbf{r}\right)$, and the bulk solvent $\tilde{A}_0\left(\mathbf{r}\right)$ that is not in the hydration shells:

$$\tilde{A}\left(\mathbf{r}\right) = \tilde{A}_1\left(\mathbf{r}\right) + \tilde{A}_0\left(\mathbf{r}\right) \tag{3.20}$$

The intensity is the Fourier transform of correlations in this electron density (where $\langle\rangle$ denotes an ensemble average):

$$\left\langle |A(\mathbf{q})|^2 \right\rangle = \int \Big[\left\langle \tilde{A}_0(\mathbf{r}) \tilde{A}_0(\mathbf{r}') \right\rangle + \left\langle \tilde{A}_1(\mathbf{r}) \tilde{A}_1(\mathbf{r}') \right\rangle + \left\langle \tilde{A}_1(\mathbf{r}) \tilde{A}_0(\mathbf{r}') \right\rangle \\ + \left\langle \tilde{A}_0(\mathbf{r}) \tilde{A}_1(\mathbf{r}') \right\rangle \Big] e^{-i\mathbf{q}\cdot(\mathbf{r}-\mathbf{r}')} d\mathbf{r} d\mathbf{r}' \tag{3.21}$$

In the "blank", we separate $\tilde{B}(\mathbf{r})$ into contribution of the water droplet $\tilde{B}_1(\mathbf{r})$ and the rest $\tilde{B}_0(\mathbf{r})$ (where the water droplet is all the water within the grid where $\tilde{A}_1(\mathbf{r})$ is non-zero), and thus similarly to Eq. 3.20 we have:

$$\tilde{B}(\mathbf{r}) = \tilde{B}_1(\mathbf{r}) + \tilde{B}_0(\mathbf{r}) \tag{3.22}$$

and an equation for $\left\langle |B(\mathbf{q})|^2 \right\rangle$ analogous to Eq. 3.21.

For the bulk solvent regions, we can write:

$$\left\langle \tilde{A}_0(\mathbf{r}) \right\rangle = \left\langle \tilde{B}_0(\mathbf{r}) \right\rangle \tag{3.23}$$

$$\left\langle \tilde{A}_0(\mathbf{r}) \tilde{A}_0(\mathbf{r}') \right\rangle = \left\langle \tilde{B}_0(\mathbf{r}) \tilde{B}_0(\mathbf{r}') \right\rangle \tag{3.24}$$

Write the cross term as:

$$\left\langle \tilde{A}_1(\mathbf{r}) \tilde{A}_0(\mathbf{r}') \right\rangle = \left\langle \tilde{A}_1(\mathbf{r}) \right\rangle \left\langle \tilde{A}_0(\mathbf{r}') \right\rangle + \alpha(\mathbf{r}, \mathbf{r}') \tag{3.25}$$

where $\alpha(\mathbf{r}, \mathbf{r}')$ is the correlation between these two points **r** and **r'**. Similarly for pure solvent system:

$$\left\langle \tilde{B}_1(\mathbf{r}) \tilde{B}_0(\mathbf{r}') \right\rangle = \left\langle \tilde{B}_1(\mathbf{r}) \right\rangle \left\langle \tilde{B}_0(\mathbf{r}') \right\rangle + \beta(\mathbf{r}, \mathbf{r}') \tag{3.26}$$

With a big enough hydration shell, we can set $\alpha(\mathbf{r}, \mathbf{r}') = \beta(\mathbf{r}, \mathbf{r}')$, since the solvent in the A_0 or B_0 region will be far from the solute and is little perturbed by it.

The intensity is now computed as the difference between a sample containing the solvent and the corresponding region in the pure solvent:

$$I(q) = \left\langle |A(\mathbf{q})|^2 \right\rangle - \left\langle |B(\mathbf{q})|^2 \right\rangle \tag{3.27}$$

Substituting Eqs. 3.21, 3.25 and 3.26 into Eq. 3.27, and using the fact that $\alpha(\mathbf{r}, \mathbf{r}') = \beta(\mathbf{r}, \mathbf{r}')$ yields:

$$I\left(\mathbf{q}\right)=\int\left[\left\langle\tilde{A}_1\left(\mathbf{r}\right)\tilde{A}_1\left(\mathbf{r}'\right)\right\rangle-\left\langle\tilde{B}_1\left(\mathbf{r}\right)\tilde{B}_1\left(\mathbf{r}'\right)\right\rangle+\left\langle\tilde{A}_1\left(\mathbf{r}\right)\right\rangle\left\langle\tilde{A}_0\left(\mathbf{r}'\right)\right\rangle+\left\langle\tilde{A}_0\left(\mathbf{r}\right)\right\rangle\left\langle\tilde{A}_1\left(\mathbf{r}'\right)\right\rangle\right.$$
$$\left.-\left\langle\tilde{B}_1\left(\mathbf{r}\right)\right\rangle\left\langle\tilde{B}_0\left(\mathbf{r}'\right)\right\rangle-\left\langle\tilde{B}_0\left(\mathbf{r}\right)\right\rangle\left\langle\tilde{B}_1\left(\mathbf{r}'\right)\right\rangle\right]e^{-i\mathbf{q}\cdot(\mathbf{r}-\mathbf{r}')}d\mathbf{r}d\mathbf{r}' \tag{3.28}$$

or

$$\begin{aligned}I\left(\mathbf{q}\right)=&\left[\left\langle A_1\left(\mathbf{q}\right)A_1^*\left(\mathbf{q}\right)\right\rangle-\left\langle B_1\left(\mathbf{q}\right)B_1^*\left(\mathbf{q}\right)\right\rangle\right]\\&+\left[\left\langle A_1\left(\mathbf{q}\right)\right\rangle-\left\langle B_1\left(\mathbf{q}\right)\right\rangle\right]\left\langle B_0^*\left(\mathbf{q}\right)\right\rangle+\left\langle B_0\left(\mathbf{q}\right)\right\rangle\left[\left\langle A_1^*\left(\mathbf{q}\right)\right\rangle-\left\langle B_1^*\left(\mathbf{q}\right)\right\rangle\right]\end{aligned} \tag{3.29}$$

which is Eq. 18 in Park *et al.* [11] From Eq. 3.22 we have

$$\left\langle B_0\left(\mathbf{q}\right)\right\rangle=\left\langle B\left(\mathbf{q}\right)\right\rangle-\left\langle B_1\left(\mathbf{q}\right)\right\rangle \tag{3.30}$$

where $\left\langle B\left(\mathbf{q}\right)\right\rangle=\int\left\langle\tilde{B}\left(\mathbf{r}\right)\right\rangle e^{-i\mathbf{q}\cdot\mathbf{r}}d\mathbf{r}$ is the Fourier transform of the shape of the entire scattering volume. In MD simulation and RISM calculation, this volume reaches infinity and thus $\left\langle B\left(\mathbf{q}\right)\right\rangle=0$ everywhere except at $q=0$, where its value is the number of electrons in that volume. The $q=0$ point is regarded as a singularity. To make our scattering curve continuous at q=0, we assume $\left\langle B\left(0\right)\right\rangle=0$, too. Hence Eq. 3.30 can be rewritten:

$$\left\langle B_0\left(\mathbf{q}\right)\right\rangle=-\left\langle B_1\left(\mathbf{q}\right)\right\rangle \tag{3.31}$$

which is essentially the Babinet's principle. Substitute that into Eq. 3.29 we have:

$$\begin{aligned}I\left(\mathbf{q}\right)=&\left[\left\langle A_1\left(\mathbf{q}\right)A_1^*\left(\mathbf{q}\right)\right\rangle-\left\langle B_1\left(\mathbf{q}\right)B_1^*\left(\mathbf{q}\right)\right\rangle\right]\\&-\left[\left\langle A_1\left(\mathbf{q}\right)\right\rangle-\left\langle B_1\left(\mathbf{q}\right)\right\rangle\right]\left\langle B_1^*\left(\mathbf{q}\right)\right\rangle-\left\langle B_1\left(\mathbf{q}\right)\right\rangle\left[\left\langle A_1^*\left(\mathbf{q}\right)\right\rangle-\left\langle B_1^*\left(\mathbf{q}\right)\right\rangle\right]\end{aligned} \tag{3.32}$$

From this we obtain a working formula for the total intensity:

$$I\left(\mathbf{q}\right)=\left[\left\langle A_1\left(\mathbf{q}\right)\right\rangle-\left\langle B_1\left(\mathbf{q}\right)\right\rangle\right]^2+\left[\left\langle\left|A_1\left(\mathbf{q}\right)\right|^2\right\rangle-\left|\left\langle A_1\left(\mathbf{q}\right)\right\rangle\right|^2\right]-\left[\left\langle\left|B_1\left(\mathbf{q}\right)\right|^2\right\rangle-\left|\left\langle B_1\left(\mathbf{q}\right)\right\rangle\right|^2\right] \tag{3.33}$$

In RISM, only the ensemble-averaged distribution of water around the solute is obtained and there is no information about the time-dependent fluctuations of $A_1(\mathbf{q})$ and $B_1(\mathbf{q})$, so that the second and third terms are not accounted for by the RISM theory. For RISM–SAXS calculation, we use the following formula:

$$I\left(\mathbf{q}\right)=\left[\left\langle A_1\left(\mathbf{q}\right)\right\rangle-\left\langle B_1\left(\mathbf{q}\right)\right\rangle\right]^2 \tag{3.34}$$

3.3.2.2 Computation of SAXS profiles

Above we show that we can approximately calculate SAXS profiles by:

$$I\left(\mathbf{q}\right) \simeq \left[\left\langle A_1\left(\mathbf{q}\right)\right\rangle - \left\langle B_1\left(\mathbf{q}\right)\right\rangle\right]^2 \tag{3.35}$$

where $A_1\left(\mathbf{q}\right)$ and $B_1\left(\mathbf{q}\right)$ are Fourier transforms for the sample and blank, respectively, but here only considering regions where there is excess/deficit electron density relative to the bulk. This approach has several considerable computational advantages for grid-based representations of the solvent. First, we do not need to include the bulk into the calculation as in Eq. 3.27, so that the three-dimensional grid need only cover regions where the solvent is perturbed by the solute. Although we only consider local regions around the solute, the fact that we compute the difference between the two amplitudes as in Eq. 3.35 effectively treats an infinitely large system (in the bulk region, $A = B$) and therefore does not introduce any artificial boundary whose shape could influence the result. Second, the amplitude computation (as in Eqs. 3.37 and 3.38 below) is linear in the number of grid points, whereas an intensity calculation (*e.g.* from a Debye sum) is quadratic in the number of grid points. (A recent study by Berlin *et al.* [157] describes an alternative approach to the Debye sum which scales with $O\left(N \log N\right)$, which might also be adapted to the grid representation used here.) The excess intensity calculation consists of two major steps, which are outlined in the following sections.

Computing the excess amplitude

We first compute the excess amplitude of the system (equivalent to $\left\langle A_1\left(\mathbf{q}\right)\right\rangle - \left\langle B_1\left(\mathbf{q}\right)\right\rangle$ in the context of Eq. 3.35):

$$\begin{aligned} A_1\left(\mathbf{q}\right) - B_1\left(\mathbf{q}\right) &= F\left(\mathbf{q}\right) \\ &= F_{solu}\left(\mathbf{q}\right) + F_{grid}\left(\mathbf{q}\right) \end{aligned} \tag{3.36}$$

where

$$F_{solu}\left(\mathbf{q}\right) = \sum_j f_j\left(q\right) \exp\left(-\frac{B_j q^2}{16\pi^2}\right) \exp\left(-i\mathbf{q}.\mathbf{r_j}\right) \tag{3.37}$$

is the form factor of the solute. The Debye–Waller factor B_j roughly accounts for thermal motion, discussed in more details in Section 3.4.1.1.

The contribution from the solvent $F_{grid}(\mathbf{q})$ is computed by performing a 3D Fourier transformation of the excess electron density, using the so-called CUBE method: [137,152]

$$F_{grid}\left(\mathbf{q}\right) = \sum_{j}^{N_{grid}} f_j\left(\mathbf{q}\right) e^{-i\mathbf{q}.\mathbf{r_j}} \tag{3.38}$$

$$f_j\left(\mathbf{q}\right) = 8\frac{\sin\left(\frac{q_x a}{2}\right)\sin\left(\frac{q_y b}{2}\right)\sin\left(\frac{q_z c}{2}\right)}{q_x q_y q_z}\rho_{xe}^{(j)} \tag{3.39}$$

where $\rho_{xe}^{(j)}$ is the excess electron density in the j$^{\mathrm{th}}$ cell of the rectangular grid, with length, width and height a, b, c. $\rho_{xe}^{(j)}$, in turn, is calculated by summing up all excess densities from individual atom types in the solution (for instance, $\mathrm{H_{wat}}$, $\mathrm{O_{wat}}$, $\mathrm{Na^+}$ and $\mathrm{Cl^-}$ in NaCl solution):

$$\rho_{xe}^{(j)} = \sum_k Z_k \rho_k \left[g_k\left(\mathbf{r_j}\right) - 1\right] \tag{3.40}$$

with ρ_k and Z_k as the bulk density and the atomic number of the k^{th} atom or ion, respectively. At each grid point the excess density is accounted for by the term $\left[g\left(\mathbf{r}\right) - 1\right]$ from the solution of the 3D-RISM equations. In Eq. 3.40, all the electrons of an atom/ion from the grid are assumed to reside within the cell where the nucleus is, in our case, a cube of length 0.5 Å. (We show in Section 3.4.1.1 below that a more realistic assignment of electron density to the grid has a negligible effect on the computed profiles.)

Computing the X-ray intensity

We next compute the excess intensity by performing spherical averaging:

$$I\left(q\right) = \frac{1}{4\pi}\int \left|F\left(\mathbf{q}\right)\right|^2 d\Omega \tag{3.41}$$

One of the fastest and most accurate ways to perform spherical integration is to use Lebedev quadrature, which is analogous to Gaussian quadrature in a linear dimension: [158,159]

$$\int \left|F\left(\mathbf{q}\right)\right|^2 d\Omega \approx \sum_{i}^{N_p} w_i \left|F\left(\mathbf{q_i}\right)\right|^2 \tag{3.42}$$

where the points are at pre-defined directions in a unit sphere (forming a 2-dimensional grid on the sphere surface) with the weights w_i. Since $I\left(\mathbf{q}\right) = I\left(-\mathbf{q}\right)$ we gain additional speed-up by evaluating the scattering vectors in only one hemisphere. As q increases, more points are needed to estimate the integral with high accuracy. For example we use 38 grid points at $q = 0.01$ Å^{-1} and 1202 grid points at $q = 1.00$ Å^{-1} which is sufficient to keep the relative errors within 10^{-3} (data not shown).

3.3.3 Computational details

We take two proteins – lysozyme and myoglobin – and different 25-bp duplex nucleic acid structures as test cases for validating the RISM–SAXS method. (Additional tests for *BioIsis.net* database are reported in Table 3.1.) The coordinates for the proteins are taken from Protein Data Bank with PDB ID 1WLA and 6LYZ for Myo and Lys, respectively. The nucleic acid starting structures are all built by the web server w3DNA [160] The DNA structure is assumed to be in B and B'-form, while the RNA is built in A-form. A-tract initial structure is taken from the server (poly d(A):poly d(T) in Na salt). Since hybrid DNA:RNA in general is known to adopt an intermediate structure between the A and B-form in solution, [161–164] we build both forms as initial structures and ran two separate MD simulations. The DNA and RNA sequence are GCAXCXGGGCXAXAAAAGGGCGXCG (where X=T for DNA and U for RNA).

3.3.3.1 MD simulation

For explicit solvent simulations, the all-atom Amber force field ff99-bsc0 was used. [165] Additional corrections for dihedral angles ε/ζ *OL1* [166] and χ *OL4* [167] were employed for DNA, and χ *OL3* [168] was used for RNA. Those corrections were found to improve the quality of the structure from MD simulation in comparison with NMR data. Since the choice of water model has been known to have a moderate effect on nucleic acid simulations, [169–171] different water models were employed here: SPC/E, [172] TIP3P, [173] TIP4PEW [174] and OPC [53]. Monovalent ion parameters were taken from Joung–Cheatham ion model. [175] We also ran simulations using CHARMM36 force field with TIP3P water. [176,177] `Parmed` in *AmberTools* was employed to convert CHARMM parameter and topology files into Amber formats. All simulations were performed using the GPU accelerated pmemd code (`pmemd.cuda`). [178–180]

Each nucleic acid was immersed in a preequilirated cubic water box with a buffer distance of 20 Å. Na^+ and Cl^- ions were added to neutralize the negative charges from the nucleic acids and reach the concentration of 100mM. Nonbonded interaction cutoff was set at 9.0 Å. The long-range electrostatic interaction is calculated by using the smooth particle mesh Ewald method. [181, 182] Equations of motion were integrated by employing the leap-frog Verlet algorithm with a 2 fs time step. Covalent bond lengths involving hydrogen atoms were constrained using SHAKE. [183] The system was first minimized with 2000 steps of steepest descent, followed by 3000 steps of conjugate gradient method to remove bad contacts. The equilibration was then performed at 298.15 K with successive solute atom restraints of 10.0, 1.0 and 0.1 kcal/(mol.Å^2) for a total of 20 ns. Temperature was regulated by using Langevin thermostat with a collision frequency of 2.0 ps^{-1} while pressure was maintained using Berendsen barostat. The production run was subsequently carried out without any restraints using Langevin thermostat and Monte Carlo barostat for each system for a total of 1 μs.

Implicit simulations were also carried out by using the GB-neck2 for nucleic acid model. [184] It was shown to give improvement results for nucleic acid simulation compared with the old GB-neck. Infinite cutoff was utilized. Salt concentration was specified at 100mM. The temperature was maintained at 298.15 K by the Langevin thermostat with a collision frequency of 1.0 ps^{-1}. For each system, dynamics were propagated for 500ns and then subjected to the clustering analysis.

Clustering analysis

Clustering analysis was performed on each trajectory using `cpptraj` in *AmberTools* 15. In brief, snapshots at every 40 ps were recorded with water molecules and ions stripped out. Clustering was performed on all atoms of the nucleic acids using the hierarchical agglomerative technique to extract 20 most representative clusters. The 20 centroid structures were subsequently subjected to 3D-RISM and SAXS calculations. The resulting intensities were then re-weighted based on each cluster size to create a single SAXS curve for each system.

3.3.3.2 RISM calculation

All calculations are performed using the `rism1d` and `rism3d.snglpnt` codes from *AmberTools*. [119] We use Amber ff12SB force field for describing the proteins and ff99-bsc0 for DNA. (Since there is no histidine-bound heme group parameter for Amber force field, we use the cysteine-bound heme parameter for Cytochrome P450 taken from Shahrokh et al. [185]) Monovalent ions (alkali, halide) parameters are taken from Joung–Cheatham ion model. [175] Sr^{2+} ion is taken from Li *et al.* (we here report the IOD set results as we find that there are no difference between SAXS calculation using these three sets). [186] The water model used in this study is cSPC/E; [119] we also did some calculations on cTIP3P water, but found that the SAXS profiles are not sensitive such a change. [119] First the 1D-RISM is carried out with only the solvent (water + ion if any) to obtain the solvent susceptibility $\chi_{\alpha\beta}^{VV}$ which contains all the information about the bulk solvent. This will be subsequently used for 3D-RISM to compute the solvent structure around a solute of choice. Thus one needs to perform only one 1D-RISM step, and use the resulting χ^{VV} for all subsequent 3D-RISM calculations which are at the same condition (salt concentration, temperature, pressure, etc). The output from RISM program is $g(\mathbf{r})$ for each atomic sites in solvents (for instance, Hw and Ow in water). These distribution functions reflect the excess or deficit of each solvent site relative to bulk concentration around the solute in real space, and can be directly used to compute SAXS profiles.

The modified direct inversion of the iterative subspace solver (MDIIS) [187] was used to iteratively solve the RISM equations to a residual tolerance of 10^{-12} and 10^{-5} for 1D and 3D-RISM, respectively at 298.15K. For 1D-RISM, the 0.025 Å grid spacing is used with 16,384 and 32,768 grid points for pure water and 100mM NaCl solution, respectively. With more diluted solutions (10mM $SrCl_2$ for example), the grid points are doubled until we get the results converged. For 3D-RISM, a 3D grid with 0.5 Å grid spacing is used with the buffer region of 20 Å for proteins and 40 Å for DNA in 100mM NaCl and up to 80 Å for 10mM $SrCl_2$.

3.3.3.3 SAXS calculation

The output from 3D-RISM program is 3-dimensional $g(\mathbf{r})$ for each atomic site in solvents (for instance, Hw and Ow in water), reflecting the excess or deficit of each solvent site relative to bulk concentration around the solute in real space. Those are then served as inputs to compute SAXS profiles using our `saxs_rism` code in *AmberTools*.

For simple and neutral protein in water (such as lysozyme, myoglobin), the RISM calculation takes 13 secs (using 1.0 Å grid spacing, 20 Å buffer), SAXS takes ~ 5 mins ($q = 1$ Å^{-1}) on a conventional desktop. For complex system (DNA in NaCl/water), a much bigger and finer box (0.5 Å grid spacing, buffer 40 Å) is required to obtain good ion distribution, RISM takes 20 mins and SAXS takes additional ~ 1 hour using 16 CPU cores.

3.4 Results and Discussions

3.4.1 Protein test cases

3.4.1.1 Lysozyme and Myoglobin

It has long been recognized that the solvent shell around a protein significantly impacts the shape of the measured SAXS profile. As a first test of the RISM–SAXS method which efficiently generates the solvent distribution around a specified solute, Figure 3.4 compares the calculated profiles of lysozyme and myoglobin with experiment and with MD simulation results reported earlier. [11] The results are shown in both logarithmic and linear scale to exploit the benefits of both – the log scale can show the overall shape of the curve, whereas a linear scale can show more clearly the details at intermediate angles.

It can be seen in both cases that RISM reproduces the peaks and troughs of the SAXS profiles and is on par with MD simulations up to $q \simeq 1.5$ Å^{-1}. Obtaining good results beyond that threshold with RISM is difficult as fluctuation effects emerge that depress the excess intensities (see Figure 3.5 in Section 3.4.1.1 below). The computed results are promising if one considers that RISM is an "implicit" solvent theory; however it differs from other implicit solvent programs (for example CRYSOL [135]) because RISM directly computes the solvent distribution around the solute considering only interactions between the solute and

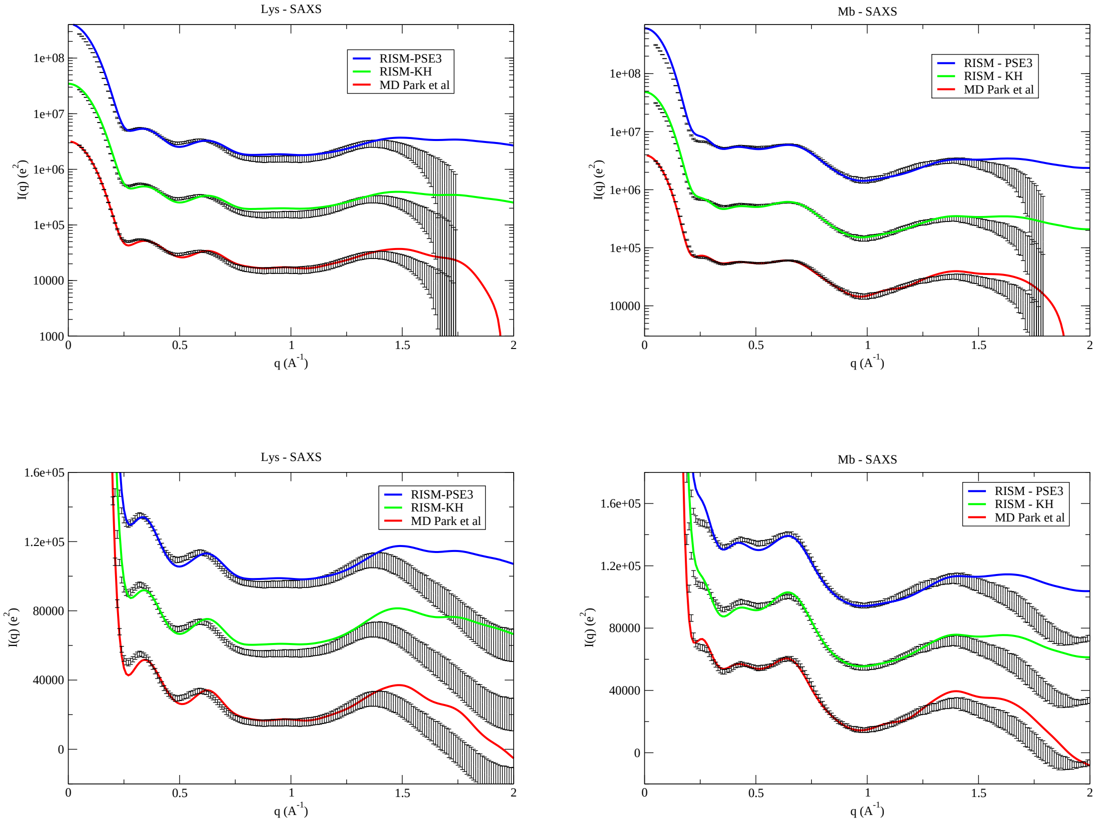


Figure 3.4: SAXS profiles of lysozyme (*left*) and myoglobin (*right*) calculated from RISM with KH and PSE3 closures, plotted in log scale (*top*) and linear scale (*bottom*). Experimental data (error bars) and MD curves (red curve) are taken from Park *et al.* [11] The data are offset for visual comparison with experiment (for the logarithmic plot, the scaling factor is 10 while for the linear plot, the offset factor is 4x10^4).

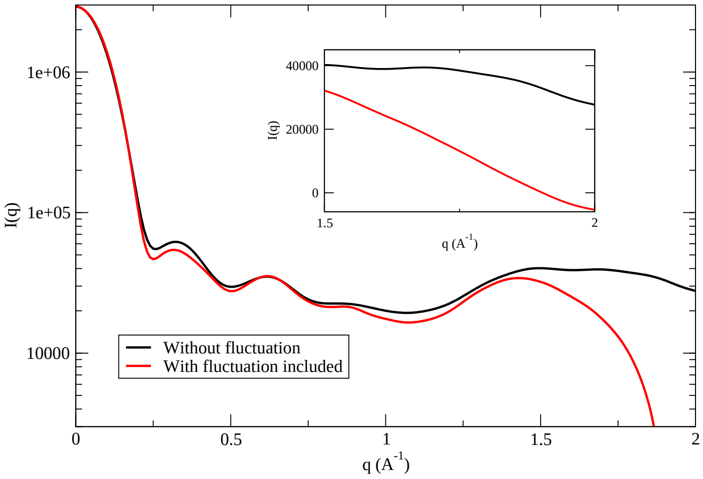


Figure 3.5: SAXS of lysozyme computed by MD simulation with (red) and without (black) the fluctuation. The inset shows two curves in linear scale at high angle.

solvent, without further assumptions or fitting parameters. The agreement between RISM–SAXS with explicit solvent models (MD) and experiment indicates that we can capture the hydration shells and SAXS curves by using RISM theory, at a fraction of the computational time associated with MD.

Effect of density fluctuations

To see the effect of the density fluctuations on the SAXS curve, Figure 3.5 illustrates the effects of ignoring the second and the third terms, using an MD simulation of Lys. It is obvious that the density fluctuations do not affect the low angle region (near $q = 0$), only moderate and high angle regions. At high angle ($q > 1.5$ Å^{-1}), the fluctuation is in the same order with the first term, even making the intensity negative (near $q = 2$ Å^{-1}).

Effect of grid fineness

We check the convergence for SAXS computation as a a function of the grid spacing in 3D-RISM (Figure 3.6). The scattering profiles at 0.5 and 1 Å grid spacing show negligible differences while at 2 Å discrepancies start showing up. Note that this is not to say scattering

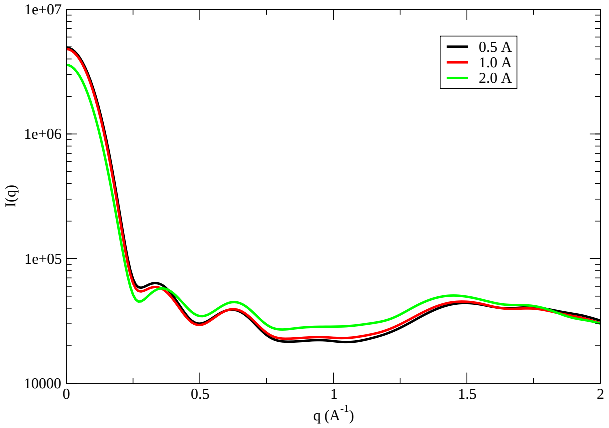


Figure 3.6: Effect of grid spacing of RISM calculation onto the scattering profile of lysozyme.

technique is able to probe the difference down to 2 Å but only show that at least a 1 Å grid spacing is needed for distribution function computed in 3D-RISM to converge.

Effect of thermal disorder

Even in the simplest case where the solute adopts a known, single and relatively rigid conformation, the scattering profiles are still affected by small thermal fluctuations of the solute. Modeling these variations as a Debye–Waller factor, as in Eq. 3.37, sometimes improves comparisons between predicted and experimental scattering profiles. [188,189] As pointed out by Moore, [190] the B-factor is only a rough guide to thermal disorder in solution, at least because B-factors are usually obtained from crystallography and are not necessarily comparable to solution scattering. In addition, correlated thermal motions (not modeled by Debye–Waller factors) contribute to scattering in solution. Fortunately, these effects are often minor, although more study is warranted. Figure 3.7 shows the effect of incorporating these effects into the calculation of Lys profiles via the B-factor as described in Eq. 3.37. Only the moderate and high angle, but not the low angle region as expected, are impacted by thermal fluctuation. The lowest scattering angle at which the thermal fluctuation effect is significant can be computed from Eq. 3.37, and is inversely related to the average B-factor. [190]

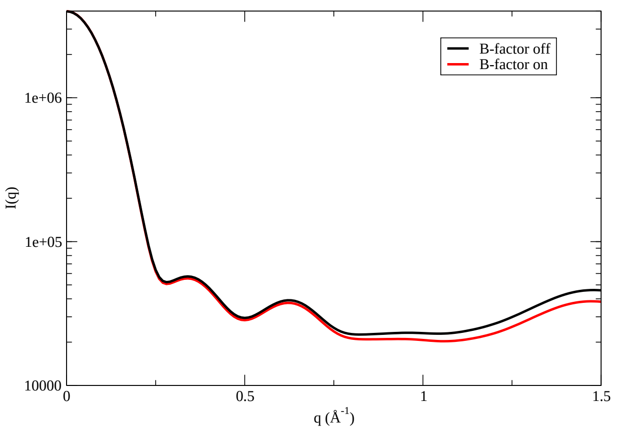


Figure 3.7: Effect of thermal fluctuation onto scattering of lysozyme. Black and red lines are scattering curves of lysozyme computed by RISM–SAXS with and without B-factor included, respectively.

Electron density calculation

The RISM model provides the density of each type of solvent component (hydrogen and oxygen of waters, plus ions) on a three-dimensional grid surrounding the solute. This needs to be converted to an electron density representation in order to compute SAXS profiles. In the simplest model (described above) atomic densities at each grid point are simply multiplied by the number of electrons in each component. In fact, the distribution of electrons around the nucleus spreads out beyond a single cube. We performed calculations where we redistribute the electrons of water over 26 neighboring cells based on the spatial distribution of electron around a single water molecule, and find the SAXS profiles are not changed up to $q = 2.0$ Å^{-1} (data not shown). This is expected since at even high angle region, the resolution in real space is still not fine enough to look at electron in atom (for example $q = 2.0$ Å^{-1} corresponds to $r \approx 3$ Å).

3.4.1.2 Comparison with other methods

Figure 3.8 shows SAXS profiles calculated by some widely used tools (CRYSOL, [135] FoXS, [146] AXES, [148] AquaSAXS [136] and HyPred [137]). For lysozyme, all of them do relatively well at small angle. CRYSOL, despite its simplicity, is able to provide an excellent

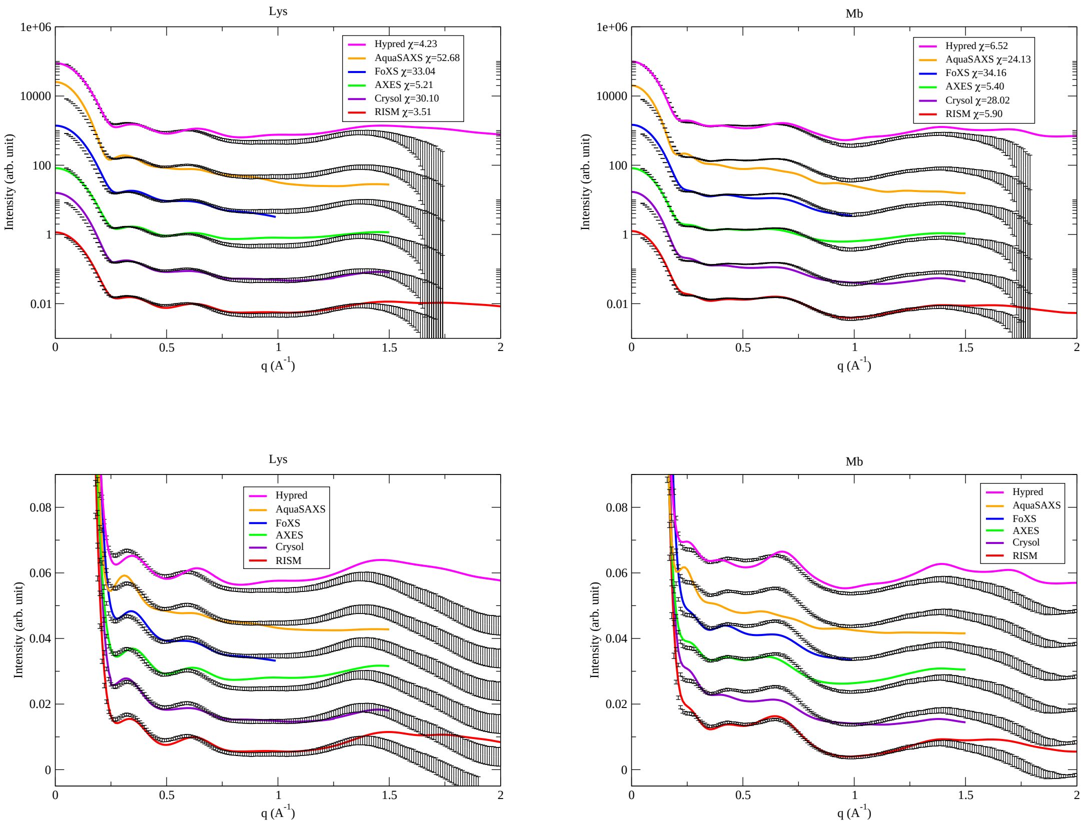


Figure 3.8: Comparison between other methods for calculation SAXS of lysozyme (left) and myoglobin (right): CRYSOL, FoXS, AXES, AquaSAXS and Hypred, plotted in log scale (top) and linear scale (bottom). The data are offset to facilitate visual comparison with experiment.

fit with the experiment up to 1.5 Å^{-1}, but overestimates scattering near $q = 0$ region. For myoglobin, no tools could predict the scattering curve satisfactorily, even at small angle region, except RISM–SAXS and, less satisfactorily, HyPred and AXES. It also should be noted that, a scaling factor is needed to plot the predicted profiles from other tools in order to match with the experiment whereas nothing similar needed in RISM–SAXS.

We have also performed calculations with structures that have experimental SAXS curves in the *BioIsis.net* database. A measure of the discrepancy between the experimental and predicted profiles is computed as:

$$\chi^2 = \sum_i \left[\frac{I_{exp}(q_i) - aI_{cal}(q_i)}{\sigma_{exp}(q_i)} \right]^2 \tag{3.43}$$

with a is the scaling constant and σ is the experimental uncertainty. As discussed above, we do not need any adjustable parameter to fit to experiments that have an absolute calibration (say against pure water); however the experimental curves in *BioIsis.net* database are all relative, and a scaling factor is needed.

Table 3.1 reports χ values between the predicted and experimental SAXS curves for several widely used tools for predicting scattering profiles with comparison with RISM. The default parameters in all these tools are used (no fitting attempt has been made). The HyPred server is not able to process nucleic acid pdb files (28BPDD and 2SAMRR), so those are not included in the statistics.

The RISM performance is encouraging, showing the best results for all tools tested with the average $\chi = 5.92$. However, it does encounter some difficulties, for instance in glucose isomerase ($\chi = 13.82$) and superoxide dismutase ($\chi = 7.69$). There are also structures that have highly flexible loops extending away from the protein core (Human regulator of chromosome condensation and glycosyl hydrolase+C-terminus), and thus are impossible to fit to experiment using only a single conformation. Whenever RISM fails, other tools also do. The model imposes a computational cost as discussed above; computation usually takes several minutes for small molecules to half an hour for biomolecules on a conventional desktop. This is faster than MD, but slower than most competing methods, and requires a force field representation before the integral equations are solved. Further study is needed to optimize this approach, especially in the presence of conformational heterogeneity.

Molecule	PDB	BioIsis–ID	q_{max}(Å)	CRYSOL	AXES	FoXS	AquaSAXS	Hypred	RISM
28 bp DNA	-	28BPDD	0.33	1.11	2.19	1.66	1.42	-	0.73
Immunoglobulin-like domains 1 and 2 of the protein tyrosine phosphatase LAR3	3PXJ	LAR12P	0.33	4.52	8.05	3.82	24.30	3.72	3.53
S-adenosylmethionine riboswitch mRNA	2GIS	2SAMRR	0.30	2.52	2.53	2.10	9.22	-	2.12
Superoxide dismutase	1HL5	APSODP	0.62	16.21	7.43	30.71	28.34	14.27	7.69
Abscisic acid receptor PYR1	3K3K	1PYR1P	0.33	7.89	8.95	3.24	28.54	16.37	4.48
Glycosyl hydrolase + C-terminus	1EDG	AT5GHP	0.60	21.12	19.70	20.01	30.25	31.06	19.41
Ubiquitin-like modifier-activating enzyme ATG7 C-terminal domain	3T7E	ATG7CP	0.33	3.04	7.88	5.56	7.02	3.29	2.64
DNA double-strand break repair protein MRE11 + ATP	3AV0	MRERAP	0.33	2.00	20.44	5.10	9.34	26.77	4.26
Glucose isomerase	2G4J	GISRUP	0.56	16.95	46.43	36.75	26.35	78.30	13.82
Complement C3b + Efb-C	-	C3BEFP	0.33	4.02	21.70	5.74	-	2.47	3.44
Xylanase	1REF	1XYNTP	0.31	3.51	3.42	4.13	6.44	1.73	1.20
Pyrococcus furiosus decameric product	2E2G	1AHRHP	0.31	5.29	7.51	5.50	6.81	6.14	5.32
Ketoreductase-enoylreductase didomain	-	ZGDWKP	0.31	2.76	4.43	4.41	5.93	3.92	2.82
Human Regulator of Chromosome Condensation	-	HRCC1P	0.33	9.92	15.68	16.05	9.97	2.72	11.37
Average				7.20	12.60	10.34	14.92	15.90	5.92
Standard deviation				6.42	11.83	11.24	10.69	22.04	5.39
Median				4.27	7.97	5.30	9.34	5.03	3.90

Table 3.1: Performance comparison (χ value) for RISM–SAXS and other tools. Structures and experimental SAXS are all taken from the *BioIsis.net* database.

3.4.2 Duplex DNA in salt solution

3.4.2.1 The ion atmosphere around duplex DNA

Few experimental techniques can directly probe the spatial distribution of ions in the positively charged cloud around DNA. [7, 191] Popular theoretical models include counter-ion condensation [84, 85] or Poisson–Boltzmann (PB) theories, [192, 193] although PB results for duplex DNA are in poor agreement with results from dialysis-type experiments. [9] MD simulation can also provide atomic details and in principle can describe the ionic atmosphere with great accuracy. [170, 194, 195] However, the computational cost is large, especially for the sampling of ions. Recently, RISM has been employed as a relatively cheap method to obtain a picture of the ion atmosphere around DNA, comparable to that of MD simulations. [9, 134]

Figure 3.9 shows an experimentally acquired SAXS profile of a 25-bp duplex DNA in 100 mM NaCl solution. This mixed sequence duplex is expected to assume the B-form. Also shown are different scattering profiles computed by RISM, including the two helical forms of DNA that best resemble the data: B and B' forms. All models are built with the w3DNA web server [160] (see Figure 3.10). The B'-form has a slightly wider major groove and narrower minor groove (the all heavy atom, *i.e.* without hydrogens, RMSD between these two structures is only 0.71 Å). This figure demonstrates the sensitivity of SAXS to the helical topology of the DNA, as mentioned above. The scattering from the B-form is in better agreement with experiment at the lowest and highest ($q > 0.6$ Å^{-1}) angles, but varies around $q = 0.4$ Å^{-1}. To further emphasize the difference in scattering profiles between these two forms at high angle, a Kratky plot of $I(q)\,q^2$ vs q is shown in Figure 3.11. This presentation of the data emphasizes the shape of the scattering profiles at larger values of q, and suggests that the B-form is generally a better representation of the duplex DNA in solution. One possible explanation for this discrepancy is that the real structure is somewhere between B and B'-form. Another possible explanation for the deviation is that uncertainties in the computed distribution of ions and water in the solution affect the result.

The important contributions of both water and ion atmospheres to the overall scattering profiles of nucleic acids are even more pronounced than solvent effects in protein systems.

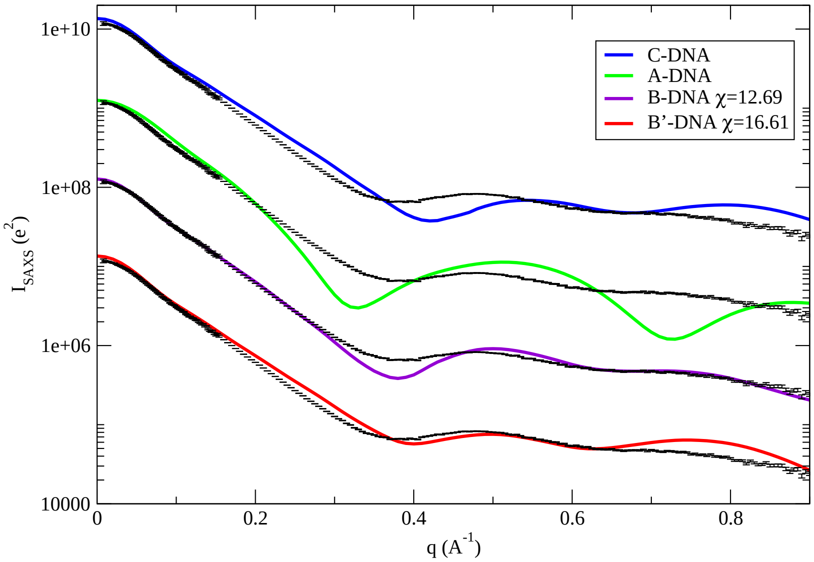


Figure 3.9: SAXS profiles of different DNA structures (built with w3DNA) computed by RISM–SAXS in 0.1 M NaCl. All the curves are offset with the scaling factor of 10 for easy comparison with the experiment (shown in error bars).

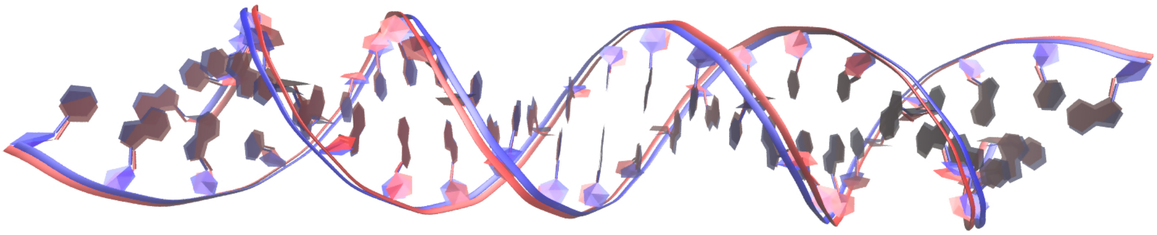


Figure 3.10: B (blue) and B'-form (red) of the 25bp duplex DNA in the solution. The differences between these two structures are trivial, with the B'-form having a slightly larger major groove and smaller minor groove. RMSD for all heavy atom = 0.71 Å, backbone only = 0.86 Å.

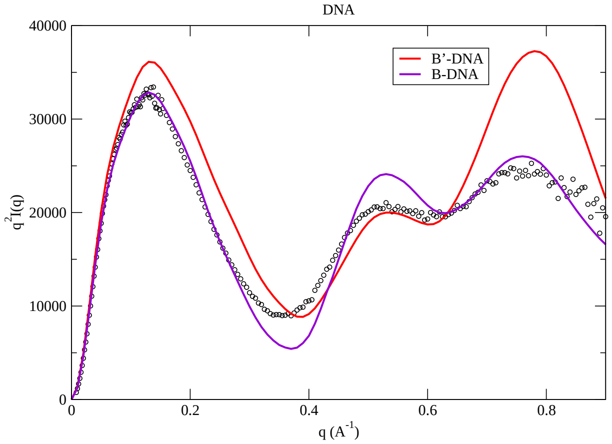


Figure 3.11: Kratky plot comparison between B and B'-form of DNA.

To demonstrate the need to properly treat the solvent in computing SAXS profiles of nucleic acids, Figure 3.12 shows the important differences in the SAXS curves that arise from the interaction of DNA with solvent. The black curve shows a SAXS profile computed from DNA atoms *in vacuo*. The dashed blue curve represents the scattering of the DNA in water, accounting only for the displacement of water by the DNA duplex, and assuming that the water molecules around it do not feel its presence and behave like bulk water. The red curve includes all contributions from the hydration shell and ion layer to the DNA scattering, and should be the most realistic calculation. Water consistently perturbs the total curve up to very high angle. The most significant changes in the scattering profiles at mid to high angle reflect both the solute topology and the behavior of the hydration layer, and underscore the sensitivity of SAXS to these different aspects of nucleic acid structure. Thus, as suggested above, discrepancies between computed and measured profiles may be useful guides for improving the accuracy of calculations.

3.4.2.2 Scattering behavior near $q = 0$

The excess form factor (as in Eq. 3.36) is the 3D Fourier transform of the excess electron distribution. At $q = 0$ it is nothing but the number of excess electron in the system. Eq. 3.36 becomes:

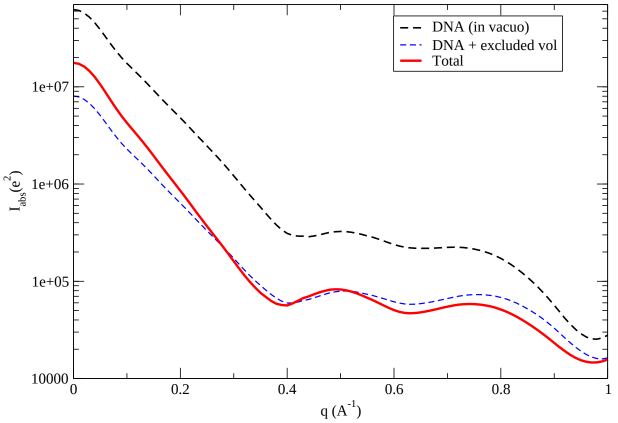


Figure 3.12: Decomposition of the total SAXS curve into contributions of the DNA+excluded volume and solvent.

$$\begin{aligned} A\left(0\right) &= F_{solu}\left(0\right) + F_{grid}\left(0\right) \\ &= \sum_i f_i\left(0\right) + \sum_k Z_k \rho_k G_k \end{aligned} \tag{3.44}$$

where G_k is the Kirkwood–Buff integral: [196]

$$\begin{aligned} G_k &= \int \left[g_k\left(\mathbf{r}\right) - 1\right] d\mathbf{r} \\ &= \int h_k\left(\mathbf{r}\right) d\mathbf{r} \end{aligned} \tag{3.45}$$

and $N_k = \rho_k G_k$ is the excess (or deficit) of the atom or ion k around the solute. Therefore the intensity at $q = 0$ is

$$I\left(0\right) = \left(Z_{solu} + Z_{wat}N_{wat} + Z_{cation}N_{cation} + Z_{anion}N_{anion}\right)^2 \tag{3.46}$$

with Z_{solu} is the number of electrons in the solute, Z_{ion} is the number in the ion, and $Z_{wat} = 10$. (Note that at $q = 0$ the form factors become real numbers.) Previous work shows that the number of excess ions extracted from RISM calculations is in good agreement with "ion counting" dialysis experiments. [9] Therefore, when absolutely calibrated SAXS data are available (see Methods), the $q = 0$ value provides an absolute comparison between data and simulation. This comparison can be used to establish the effectiveness of RISM

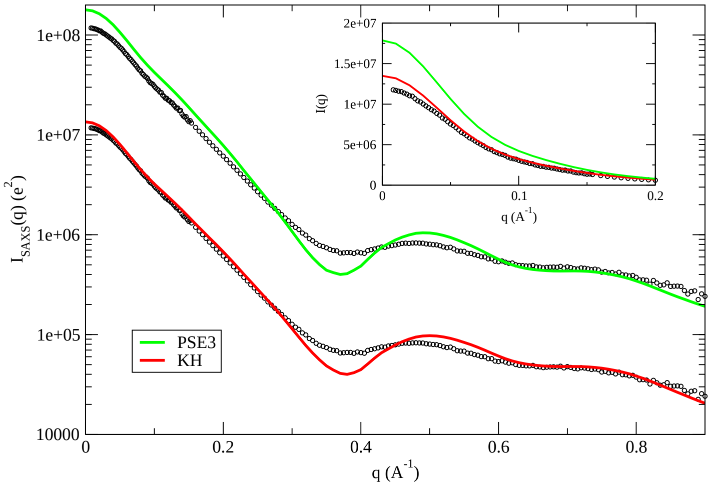


Figure 3.13: SAXS profiles of B-DNA computed by RISM coupled with KH (red) and PSE3 (green) closures. The two profiles are offset by a factor of 10 for easy comparison. The inset zooms out the low angle region near $q = 0$, and is plotted without the offset factor.

subject to different closures. Figure 3.13 compares measured and calibrated SAXS profiles of a 25-bp duplex DNA in 100 mM NaCl with profiles computed with two closures, KH and PSE3. The KH curve agrees better with the experiment near $q = 0$, implying that the total number of excess electrons in the system should be closer to those from KH as opposed to PSE3 closure.

Table 3.2 reports the number of excess waters and ions around the 25-bp DNA computed by RISM coupled with different closures. The neutral atomic form factors are used in the

	N_{wat}			N_{Na}	N_{Cl}	$N_e = \sqrt{I(0)}$
	N_{excl}	N_{shell}	Total			
KH	-592	172	-420	30.5	-17.5	3722
PSE2		196	-396	35.6	-12.4	4104
PSE3		207	-385	37.7	-10.3	4275
PSE4		211	-381	39.1	-8.9	4356
SAXS		107	-485 ± 16	39 ± 2	-9 ± 2	$3,300 \pm 100$

Table 3.2: Number of excess water and ions around the 25-bp duplex DNA from SAXS experiment and RISM calculations with various closures. N_{wat} is partitioned into contributions from the excluded volume of the DNA N_{excl} and hydration shell N_{shell}, as described in the text.

SAXS calculations, requiring a modification of the electron number N_e to account for the overall charge of the DNA: we correct the computed N_e to include the extra electrons accounting for the DNA charge. The number of excess water is approximately partitioned into contributions from the excluded volume of the DNA, and the remainder, which is termed the hydration shell. (Whether a cell belongs to the excluded volume or hydration shell depends on the distance d between it and its nearest atom of the solute j. If $d < r_j + r_{wat}$ where r_j is the atomic radii of atom j and $r_{wat} = 1.4$ Å is approximately the radius of water molecule then the cell is within the excluded volume. This is a somewhat arbitrary division, and the results in Table 3.2 provide only a general account of excluded-volume versus hydration shell effects.)

From Eq. 3.46, an experimental estimate of N_{wat} can be extracted if all other terms are known. We assume here that the number of excess Na$^+$ is the same as that measured for Rb$^+$, which is 39 ± 2 (see below), which in turn implies that N_{Cl} is -9 ± 2, to achieve electroneutrality. The total number of electrons in the DNA is 7940 (assuming a net charge of -48), and $I(0) = 1.098 \pm 0.070 \times 10^7$, extrapolated using GNOM, [197] then the number of excess waters can be computed from Eq. 3.47:

$$\begin{aligned} 1.098 \times 10^7 &= (Z_{DNA} + 10N_{wat} + 10N_{Na} + 18N_{Cl})^2 \\ &= (7940 + 10 \times N_{wat} + 10 \times 39 + 18 \times [-9])^2 \end{aligned} \tag{3.47}$$

This gives $N_{wat} = -485 \pm 16$. As shown in Table 3.2, this total can be approximately viewed as the sum of a deficit of -592 waters (arising from the excluded volume of the DNA duplex), and an excess of 107 waters in the hydration shell. All of the RISM closures appear to overestimate the number of excess waters in the hydration shell. The KH results are closest to experiment, with an overestimate of about 2 water molecules per base-pair.

3.4.2.3 Anomalous SAXS

The use of ASAXS (see Section 3.2.4) provides another important degree of comparison between RISM and measurement. The ASAXS profile is the difference between the SAXS curves of the same sample but probed at two different beam energies. One of these energies is close to the absorption edge of a particular element, in this case Rb$^+$ or Sr^{2+}. The energy

Ion	f' on-edge	f' off-edge
Rb^+	-7.02	-3.58
Sr^{2+}	-6.99	-3.61

Table 3.3: Anomalous scattering factors measured for Rb and Sr in on-edge and off-edge experiments. The imaginary factor $f'' = 0$.

change only influences the scattering of the selected ions, but the ASAXS signal contains contributions from all terms involving ion scattering, including ion-ion, ion-solute and ion-water cross terms. ASAXS experiments are restricted to elements whose K-edges are readily X-ray accessible. Past work focused on Rb, Sr and Co. Lighter and more biologically relevant elements (O, C, N, Na, Mg ...) which have low energy K-edges are currently inaccessible for ASAXS.

The atomic scattering factor for the ions probed by ASAXS can be written as Eq. 3.7:

$$f(q, E) = f_0(q) + f'(E) + if''(E) \tag{3.48}$$

where f' and f'' are the anomalous scattering factors and they do not depend on the scattering angle in SAXS. The X-ray energy beam is chosen right below the absorption edge of the ion, so that f'' is effectively zero and the imaginary part does not contribute to the change of the atomic scattering factor (see Figure 3.1). Table 3.3 reports the values of the real part f' for Rb^+ and Sr^{2+} used in ASAXS.

Figure 3.14 shows a comparison of experimental ASAXS profiles of Rb^+ and Sr^{2+} around duplex DNA with profiles computed from RISM–SAXS, shown in the absolute scale. ASAXS curves of a similar RNA sequence, also in RbCl and $SrCl_2$, were computed by MD simulation and reported earlier, however in the relative scale. [195] In contrast to comparisons on the full SAXS profiles of B-DNA, where the PSE-n closures are not as good as the KH closure in terms of matching with the experiment near the $q = 0$ region, the PSE-n closures give better results when compared with ASAXS data. Earlier work has shown that PSE-n gives better results than KH for ion distributions around nucleic acids. [9] Since the ASAXS profile is a complicated sum of ion-solute, ion-water and ion-ion terms, it is not surprising that none of the calculated ASAXS curves from RISM fit the experimental data. Due to weaker site-site interactions, KH closure places ions farther from the solute, leading to a more rapid decay of the ASAXS curve than expected from the PSE2 and PSE3 closures.

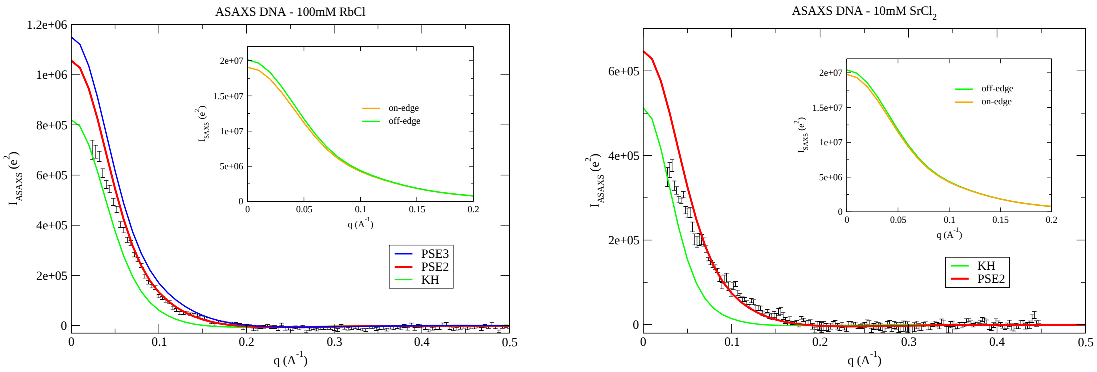


Figure 3.14: ASAXS signals of Rb^+ (left) and Sr^{2+} (right) around the 25-bp duplex DNA computed from RISM–SAXS coupled with different closures, with the experimental profiles shown as error bars. The insets show the calculated on-edge and off-edge SAXS profiles in which only the atomic scattering factors of the cations are varied. The ASAXS signal is obtained by subtracting the on-edge from the off-edge.

3.4.3 Ensemble-based SAXS of the DNA duplex

It is established in Section 3.4.2 that the intensity profiles of the DNA duplex depend strongly on the helical topology. We then perform an MD simulation with the fully flexible DNA to study the effect of solute flexibility on the scattering profiles. Snapshots from the trajectories are then subjected to the clustering analysis to pick out 20 most representative structures from the simulation (see Section 3.3.3 for more technical details). Those structures are served as the inputs for 3D-RISM and SAXS calculations. The resulting intensities are then averaged (with weights based on each cluster size) to generate a single SAXS curve. Figure 3.15 compares the scattering curve of the static B-form structure and the ensemble-averaged curve. Although there are still some discrepancies around the moderate angle region and near $q = 0$, it is clear that the ensemble-averaged intensity profile displays significant improvement versus the static structure.

3.5 Conclusions

Small angle X-ray scattering experiment, in addition to the macromolecule shape in general, provides important information into how the molecule modifies the bulk solvent. Here we describe a method using the solvent distribution from 3D-RISM model to calculate X-ray

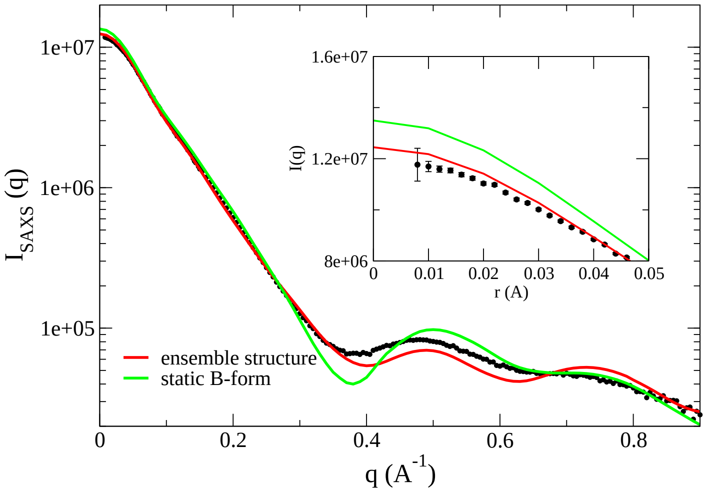


Figure 3.15: Comparison between SAXS profiles of (static) B-form duplex DNA (green) vs. ensemble structure from MD simulation (red) (plotted in logarithmic scale). The inset zooms out the small angle region and is plotted in linear scale.

scattering profiles for proteins and nucleic acids. These integral equation models are certainly not perfect, but provide usefully accurate X-ray intensity profiles that agree better with experiment for a number of test cases than do the predictions of other simple models, and rival the results of much more expensive MD simulations. 3D-RISM is particularly useful for cases where there are both ions (and cosolvents in general) and water in the environment, since there are few existing implicit models that describe both, and equilibration of ion densities in MD simulations can be prohibitively expensive to achieve (especially for diluted concentrations).

The basic analysis described here uses a single structure to describe the solute biomolecule. Even relatively rigid biomolecules may have solute conformational fluctuations that can affect scattering profiles in the wide-angle region beyond $q \approx 0.3$ Å^{-1}. A very simple approach models these fluctuations in the same way as do atomic displacement parameters (or B-factors) in crystallography. This model provides some insight, but ignores differences between the crystal and solution environment, and fails to include the effects of correlated fluctuations that affect solution scattering but not the intensities of Bragg peaks in crystallography. Averaging over snapshots from MD simulations offers one way to investigate such effects. We illustrate that for a duplex DNA, the total intensity calculated by averaging over snapshots from MD simulation is generally better the one computed with the static diffraction structure. However, more work in this area still needed. We consider here only the "forward" challenge of estimating SAXS profiles based on an input structure; the "inverse" problem of constructing a structure or ensemble consistent with a given profile is more challenging, and is generally problem-specific. Our computation is fast enough (requiring a few minutes for the examples considered here) to allow one to average over many solute configurations, or to use SAXS results (perhaps in combination with other restraints) to construct ensembles of configurations consistent with the data. [198, 199]

Analysis of the $q = 0$ limits, and comparison to experimentally calibrated profiles, allows one to count the numbers of excess ions and waters in the vicinity of biomolecules. These in turn can be used to test the accuracy of computations (including finding the limitations of the RISM model used here), and to complement other ion counting experiments. These counts are related to partial molar volumes and contributions to osmotic pressure, and offer insights into molecular interactions and function. A preliminary example, of duplex DNA in NaCl/water, suggests that the excess number of waters surrounding the duplex can be estimated with a precision of 1-2 water molecules per base pair, and that the force fields and RISM models used here tend to overestimate the number of excess waters. (This tendency only affects the scattering curves for $q < 0.05$ Å^{-1}, and generally good results are obtained for higher scattering angles.)

The characterization of the solvent perturbation used here relies on a thermally–averaged density profile, and appears to be only appropriate for $q < 1.5$ Å^{-1}. At wider angles, fluctuations in the solvent densities (and not just the average density) become important, and a different type of theory is needed. (At high angles, errors in the 3D-RISM description of pure water may also be a factor limiting the application of this model.) Nevertheless, this range of scattering angles covers a large fraction of reported scattering profiles, and our model should be of considerable use. The programs used here are incorporated into the *AmberTools* suite of programs (`rism_saxs` and `rism_md`), available at `http://ambermd.org`.

Chapter 4

Extracting water and ion distributions from SAXS

4.1 Introduction

Ions and water molecules have been long known to play crucial roles in governing biomolecule stability and function. [26–28, 67, 70, 76, 200] Elucidating how ions and water molecules distribute themselves around the solutes can provide valuable insights in their function, and can also provide experimental tests for theoretical predictions. [67, 69, 70, 76, 201, 202] However, there are few experimental methods that directly probe the positions of ions and water molecules in the solution. Ion counting via dialysis can provide a quantitative measure of the ionic atmosphere around the solute, but it provides only a total (excess) number, and not any information the shape of the ion cloud. [7, 203] The $q = 0$ limit of small-angle X-ray scattering (SAXS) can provide similar excess counts for both water and ions, [12] as discussed in Chapter 3. Anomalous small angle X-ray scattering (ASAXS) data in principle yield additional information about the extent of perturbations of the ion/water environment, [106, 191, 204, 205] but the ASAXS signal is known to be intertwined with all components in the system, complicating the analysis (see more discussion in Section 3.4.2.3). [12, 206] Efforts to extract the contribution from ions to ASAXS profiles date back to 2003 with the work of Ballauff and coworkers. [204, 207] Using multiple energy ASAXS, they were able to decompose the total scattering intensity into solute-ion and ion-ion contributions, though limited to spherical solutes. Recently, Meisburger *et al.* proposed a similar approach to separate information about the ion contribution around a DNA duplex, using the heavy ion replacement SAXS rather than ASAXS profiles. [13] Both approaches, nonetheless, only show the solute-ion and ion-ion correlation in reciprocal space and are not readily extended to the water distribution.

In this chapter, we present a new method of analysis to extract both water and ion distributions from SAXS profiles provided the scattering intensities are calibrated on the absolute scale. Both the excess number (as in ASAXS and ion-counting experiments [7,106]) and (low-resolution) information about the actual distribution can be obtained for ions and water molecules. The correlation between solute-ion or solute-water can be displayed in both reciprocal space (as partial intensities) or real space (as interatomic distribution functions). Since the focus is on ions and water, the proposed analysis requires knowledge about the biomolecule structure in advance. Although the proposed method is approximate, we use theoretical models to demonstrate that the errors are rather small for q between 0 and 0.1 Å^{-1}. The resulting ion and water distributions are then used to test predictions from integral equation theory and explicit MD simulation for relatively rigid proteins and a DNA duplex.

4.2 General theory

The calculation of SAXS profiles from atomic coordinates is discussed in details in Section 3.3. Here we briefly rewrite key equations and then transition to the derivation of the analysis scheme.

4.2.1 Calculation of SAXS profiles

X-ray scattering experiments on biomolecules compare the scattering intensity from the sample of interest to a “blank” with just solvent present, and report the difference, or “excess” intensity:

$$I\left(\mathbf{q}\right)=\left\langle \left|A\left(\mathbf{q}\right)\right|^{2}\right\rangle_{t}-\left\langle \left|B\left(\mathbf{q}\right)\right|^{2}\right\rangle_{t} \tag{4.1}$$

where the $\langle\rangle_t$ bracket indicates the intensities are averaged over the measurement time and volume. $A\left(\mathbf{q}\right)$ and $B\left(\mathbf{q}\right)$ are Fourier transforms of the scattering amplitudes for the sample and blank, respectively:

$$\left\langle \left|A\left(\mathbf{q}\right)\right|^{2}\right\rangle=\int\left\langle \tilde{A}\left(\mathbf{r}\right)\tilde{A}\left(\mathbf{r}'\right)\right\rangle e^{-i\mathbf{q}.\left(\mathbf{r}-\mathbf{r}'\right)}d\mathbf{r}d\mathbf{r}' \tag{4.2}$$

with $\tilde{A}\left(\mathbf{r}\right)$ is the electron density in the system. It has been shown that the total intensity can be approximately (though usefully) rewritten as:

$$I\left(\mathbf{q}\right) = \left[\left\langle A_1\left(\mathbf{q}\right)\right\rangle - \left\langle B_1\left(\mathbf{q}\right)\right\rangle\right]^2 + \left[\left\langle \left|A_1\left(\mathbf{q}\right)\right|^2\right\rangle - \left|\left\langle A_1\left(\mathbf{q}\right)\right\rangle\right|^2\right] - \left[\left\langle \left|B_1\left(\mathbf{q}\right)\right|^2\right\rangle - \left|\left\langle B_1\left(\mathbf{q}\right)\right\rangle\right|^2\right] \tag{4.3}$$

where $A_1\left(\mathbf{q}\right)$ and $B_1\left(\mathbf{q}\right)$ are Fourier transforms for the sample and blank, respectively, but here only considering regions where there is excess/deficit electron density relative to the bulk value. In 3D-RISM, the second and third terms vanish, leading to:

$$I\left(\mathbf{q}\right) = \left[\left\langle A_1\left(\mathbf{q}\right)\right\rangle - \left\langle B_1\left(\mathbf{q}\right)\right\rangle\right]^2 \tag{4.4}$$

The approximation made in going from Eq. 4.3 to Eq. 4.4 has been shown valid up to $q = 1.5$ Å^{-1} in Section 3.4. Finally, the angular averaging is performed to obtain the total intensity using the Lebedev quadrature:

$$I\left(q\right) = \frac{1}{4\pi}\int I\left(\mathbf{q}\right) d\Omega \tag{4.5}$$

The total excess amplitude can be expressed as the sum of terms arising from the solute (biomolecule) and the solvent:

$$\begin{aligned} A_1\left(\mathbf{q}\right) - B_1\left(\mathbf{q}\right) &\equiv F\left(\mathbf{q}\right) \\ &= F_{solu}\left(\mathbf{q}\right) + F_{solv}\left(\mathbf{q}\right) \end{aligned} \tag{4.6}$$

The separation between F_{solu} and F_{solv} can be made in different ways, and is primarily for convenience in interpreting results. Here we have chosen to include in the F_{solu} term the scattering from the excess electron density in the region of space occupied by the solute:

$$F_{solu}(\mathbf{q}) = \sum_j f_j\left(q\right)\exp\left(-B_j q^2/16\pi^2\right) e^{-i\mathbf{q}\cdot\mathbf{r_j}} + \int_{exclV} f_k\left(\mathbf{q}\right) e^{-i\mathbf{q}\cdot\mathbf{r_k}} d\mathbf{r_k} \tag{4.7}$$

The first term on the right-hand-side represents the scattering from the solute atoms *in vacuo*, where $f_j\left(q\right)$ is the atomic scattering factor and B_j is the B-factor of atom *j*. The second term gives the contribution from the (negative) excess solvent density in the volume occupied by the solute. As in earlier work, [12,137,152] we use a "cube method" to compute the scattering from a three-dimensional voxel, so that:

$$f_k\left(\mathbf{q}\right) = 8\left[\sin\left(\frac{q_x a}{2}\right)\sin\left(\frac{q_y b}{2}\right)\sin\left(\frac{q_z c}{2}\right)/\left(q_x q_y q_z\right)\right]\rho_{xe}\left(\mathbf{r_k}\right) \tag{4.8}$$

where $\rho_{xe}\left(\mathbf{r_k}\right)$ is the excess electron density arising from the solvent; *a*, *b*, *c* are the grid length, width and height, respectively, and the integral only goes over all points within the excluded volume of the solute.

With this definition $F_{solu}(\mathbf{q})$ is then the scattering amplitude of a hypothetical system where the solute displaces waters inside its volume, but does not affect the water molecules and ions around it. The details of how to determine the solute excluded volume, and the choice to include the "excluded" waters in F_{solu}, are somewhat arbitrary. Here, the excluded volume is computed based on the algorithm of Voss and Gerstein for the 3D grid with the probing radius of 1.4 Å. [16] Points lying inside this volume will be assigned to the excluded volume. A key advantage of Eq. 4.7 is that it is readily calculated for a solute of known structure: the atomic positions and scattering factors are known, and the excess solvent density inside the molecule is just the negative of the bulk solvent density. The "interesting" parts of solvation, *i.e.* how the solvent in the vicinity of the solute is perturbed, are included in $F_{solv}(\mathbf{q})$. The key point of this analysis, given in the next section, is to show how $F_{solv}(\mathbf{q})$ can be extracted from experimental data.

4.2.2 Extracting water and ion distributions from SAXS and ASAXS

We can further divide the solvent scattering into excess terms arising from water and from ions:

$$F\left(\mathbf{q}\right) = F_{solu}\left(\mathbf{q}\right) + F_{hyd}\left(\mathbf{q}\right) + F_{ion}\left(\mathbf{q}\right) \tag{4.9}$$

As above, the solvent terms come from scattering outside the biomolecule:

$$F_{hyd}\left(\mathbf{q}\right) + F_{ion}\left(\mathbf{q}\right) = \int_{not-exclV} f_k\left(\mathbf{q}\right) e^{-i\mathbf{q}.\mathbf{r_k}} d\mathbf{r_k} \tag{4.10}$$

Since $f_k\left(\mathbf{q}\right)$ is proportional to the excess electron density $\rho_{xe}\left(\mathbf{r_k}\right)$ we can further decompose the solvation shells into contributions of hydration water and ions by considering excess electron density coming from only water or ions $\rho_{xe}\left(\mathbf{r_k}\right) = \rho_{xe}^{(wat)}\left(\mathbf{r_k}\right) + \rho_{xe}^{(ion)}\left(\mathbf{r_k}\right)$. Note that this particular decomposition reflects our interest in studying the waters of hydration and ions around the solute. In principle, any decomposition scheme should work.

Now, as the total intensity is

$$I\left(q\right) = \frac{1}{4\pi} \int \left|F\left(\mathbf{q}\right)\right|^2 d\Omega \tag{4.11}$$

we define, similarly, the partial intensity for each component γ (where $\gamma =$ *solu*, *hyd, counterion* or *co-ion*):

$$I_\gamma\left(q\right) = \frac{1}{4\pi}\int\left|F_\gamma\left(\mathbf{q}\right)\right|^2 d\Omega \tag{4.12}$$

The square root of the partial intensity will be called $\tilde{F}_\gamma\left(q\right)$, by definition a real quantity. At $q = 0$ it is nothing but the number of excess electrons from component γ, *i.e.* $\tilde{F}_\gamma\left(0\right) = N_\gamma Z_\gamma$ where N_γ is the excess number of component γ coming from the hydration shell and Z_γ is the number of electrons of component γ. Thus we always have $\tilde{F}\left(0\right) = \sum_\gamma \tilde{F}_\gamma\left(0\right)$, or

$$\sqrt{I\left(0\right)} = \sum_\gamma N_\gamma Z_\gamma \tag{4.13}$$

and we can extract the number excess of particle N_γ from $I\left(0\right)$, as illustrated in the previous Chapter. At non-zero q, $F_\gamma\left(\mathbf{q}\right)$ is generally complex, but we expect there should be a small range at low angles where one can still decompose the total into partial amplitudes:

$$\tilde{F}\left(q\right) \simeq \sum_\gamma \tilde{F}_\gamma\left(q\right) \tag{4.14}$$

For the above equality to be true, the phases of $F_\gamma\left(\mathbf{q}\right)$ should be identical (or at least very close to each other) at all small angles. The phase difference α between $F_{solu}\left(\mathbf{q}\right)$ and $F_{hyd}\left(\mathbf{q}\right)$ is, by definition:

$$F_{solu}\left(\mathbf{q}\right) F_{hyd}\left(\mathbf{q}\right) = \left|F_{solu}\left(\mathbf{q}\right)\right| \left|F_{hyd}\left(\mathbf{q}\right)\right| \cos\alpha \tag{4.15}$$

Since $\tilde{F}\left(q\right)$ depends on all possible orientations of the $\boldsymbol{q}$ vector (with the same magnitude q), the approximation in Eq. 4.14 holds if and only if $\cos\alpha$ is very close to 1 ($\alpha \approx 0$) for every $\boldsymbol{q}$. As we show below, this condition is valid for q less than about 0.1 Å^{-1}. Since SAXS information content is relatively low, [208] it is not obvious that we we can capture enough “useful” information from such a narrow region to reconstruct the real-space water distribution. However we also show below that the information content in the small angle region is great enough to construct a usefully accurate approximate pair distance distribution function (PDDF) from the estimated excess amplitude.

4.2.2.1 Solvent is pure water

For systems with only the solute in pure water, the third term in Eq. 4.9 vanishes, therefore we can extract directly $\tilde{F}_{hyd}\left(q\right)$ from SAXS:

$$\tilde{F}_{hyd}\left(q\right) = \sqrt{I\left(q\right)} - \tilde{F}_{solu}\left(q\right) \tag{4.16}$$

where $\tilde{F}_{solu}(q) = \sqrt{I_{solu}(q)}$, which can be computed from the (known) structure of the solute. The standalone $\tilde{F}_{hyd}(q)$ is useful but only contains information about the water of hydration. To describe how those water molecules distribute around the solute, we seek an approximation of the cross-term *solute-hyd* which is the correlation between the hydration water density and the solute. From Eq. 4.11:

$$\begin{aligned} I(q) &= \frac{1}{4\pi}\int \left|F_{solu}(\mathbf{q}) + F_{hyd}(\mathbf{q})\right|^2 d\Omega \\ &= \frac{1}{4\pi}\int \left|F_{solu}(\mathbf{q})\right|^2 d\Omega + \frac{1}{4\pi}\int \left|F_{hyd}(\mathbf{q})\right|^2 d\Omega \\ &+ \frac{1}{4\pi}\int \left[F_{solu}(\mathbf{q}) F^*_{hyd}(\mathbf{q}) + F^*_{solu}(\mathbf{q}) F_{hyd}(\mathbf{q})\right] d\Omega \end{aligned} \tag{4.17}$$

The first two terms are $\tilde{F}^2_{solu}(q)$ and $\tilde{F}^2_{hyd}(q)$ (as defined in Eq. 4.12), respectively, therefore we can approximate the cross-term (the third term) as

$$2\tilde{F}_{solu}(q)\tilde{F}_{hyd}(q) = I(q) - \tilde{F}^2_{solu}(q) - \tilde{F}^2_{hyd}(q) \tag{4.18}$$

As before, we can compute $\tilde{F}_{solu}(q)$ from the structure of the biomolecule, $\tilde{F}_{hyd}(q)$ from Eq. 4.16, and get the cross-term from Eq. 4.18. Examples of this sort of analysis are given below.

4.2.2.2 Solvent contains ions and water

In this case, one has more than one unknown (from *hyd, counterion* and *co-ion*) in Eq. 4.9, and additional measurements or assumptions are required to carry out the decomposition. Changing the energy of the incident beam in an ASAXS experiment is one approach, varying the atomic scattering factor of a given ion. [106, 191, 204, 206, 209, 210] Another approach uses heavy ion replacement, assuming the ion and water distributions are similar for the same type of ions (alkalies, for example). [13, 191] By doing this, only $\tilde{F}_{counterion}$ is allowed to vary while the co-ion and hydration terms are fixed. Subtracting the square roots of two measured intensities, therefore, gives the contribution from the counterion only:

$$\begin{aligned} \tilde{F}(q) - \tilde{F}'(q) &= \tilde{F}_{counterion}(q) - \tilde{F}^{'}_{counterion}(q) \\ &= N_{counterion}\left(Z_{counterion} - Z^{'}_{counterion}\right)\tilde{f}_{counterion}(q) \end{aligned} \tag{4.19}$$

with Z and Z' are the atomic scattering factors $(q = 0)$ at two different energy beams, $\tilde{f}$ is the normalized $\tilde{F}$ $(\tilde{f}(0) = 1)$. This can be scaled up to compute back the "full" term:

$$\begin{aligned}\tilde{F}_{counterion}(q) &= N_{counterion} Z_{counterion} \tilde{f}_{counterion}(q) \\ &= \frac{Z_{counterion}}{Z_{counterion} - Z'_{counterion}} \left[\tilde{F}(q) - \tilde{F}'(q)\right]\end{aligned} \tag{4.20}$$

[Note that this $\tilde{F}_{counterion}(q)$ includes a small contribution from the solute excluded volume effect, since "inside" the solute there is a deficit density of the ion. This contribution, however, is small due to the small concentration of ions normally used in SAXS.] Rewriting Eq. 4.14 as:

$$\tilde{F}_{hyd}(q) + \tilde{F}_{co-ion}(q) = \tilde{F}(q) - \tilde{F}_{solu}(q) - \tilde{F}_{counterion}(q) \tag{4.21}$$

The co-ion contribution in principle could be accurately subtracted from $\tilde{F}_{hyd}(q)$ but only at $q = 0$ since we only know its number of excess thanks to the electroneutrality principle: if the total charge of the solute is Z, and the number of excess counterions $N_{counterion}$ can be computed as $\tilde{F}_{counterion}(0)/Z_{counterion}$, the number of excess co-ions must be $N_{co-ion} = N_{counterion} - Z$. However, there is currently no way to obtain the co-ion spatial distribution from the experiment. ASAXS experiments of DNA in NaBr or NaI with beam energies close to the absorption edge of Br or I could potentially provide the answer for this. One very primitive way is to construct a box around the solute and assume the co-ions are completely depleted inside this box $(g = 0)$, whereas outside this box, its concentration returns back to the bulk value $(g = 1)$. The shape of the box ideally closely resembles the solute, although it is acceptable to use a rectangular box for the DNA here. The size of the box is chosen so that the number of excess of co-ion in this box exactly matches the calculation above. With $\tilde{F}_{co-ion}(q)$ approximately determined, we can now extract the hydration term $\tilde{F}_{hyd}(q)$. The reason we need to account for the co-ion term is because its magnitude is in the same order of the counterion and hydration water terms, although the co-ion term gives negative contribution due to the depletion of the co-ions. (The importance of co-ion exclusion is emphasized in a recent "ion counting" study by Herschlag and colleagues. [211]) More sophisticated models for the co-ion distribution are possible, and are under consideration.

As for the pure water case above, it is probably most useful to compute the cross-terms $\tilde{F}_{counterion}(q)\tilde{F}_{solu}(q)$ or $\tilde{F}_{hyd}(q)\tilde{F}_{solu}(q)$. Examples are shown below.

4.2.3 Pair distance distribution function (PDDF)

4.2.3.1 Indirect Fourier transform (IFT)

The PDDF is the ultimately sought quantity and carries the most information content from SAXS experiment. In principle, a direct Fourier transform of the pseudo intensity $\left|\tilde{F}_{hyd}(q)\right|^2$ should provide a real space representation of the distribution of hydration water under the form of the PDDF $p_{hyd}(r)$. However, such an approach is of little use since it gives large systematic errors because of the noise, smearing and truncation of the experimental data. Instead, indirect Fourier transform (IFT) technique was developed long time ago, pioneering by Glatter, [212] Moore [208] and Svergun *et al.* [213] to deal with this problem. The main idea of IFT was to express the PDDF $p(r)$ as a linear combination of a set of smooth basis functions and fit the coefficients in order to reproduce the intensity in the reciprocal space. In this work, we used the IFT method based on Bayesian analysis from Hansen, which was shown to give similar results with Glatter method; all tranformations were performed using the webserver BayesApp. [214–216] Since IFT is an underdetermined problem, *i.e.* several solutions can fit the experimental data equally well, additional contraints must be introduced to obtain unique solution. [208,212–214] The PDDF is first written as a sum of smooth basis functions (cubic B-splines for instance) $p(r) = \sum_i a_i B_i(r)$. The coefficients a_i are then tuned by minimizing the regulation functional S, subjected to the constraint that χ^2 takes sensible value, with $\chi^2 = \sum \left[I_{exp}(q_j) - I_{trans}(q_j)\right]^2 / \sigma_j^2$ and σ_j is the standard deviation at data point j. The regulation functional S usually controls the smoothness of the solution and several forms of it exist. We followed Hansen and used $S = \int p''(r)^2 \, dr$. [214–216]

4.2.3.2 Direct calculation

The PDDF is related to the density-density correlation by: [217,218]

$$p(r) = \left\langle \int_V \rho(\mathbf{r}') \, \rho(\mathbf{r} + \mathbf{r}') \, d\mathbf{r}' \right\rangle \tag{4.22}$$

where $\rho(\mathbf{r})$ is the scattering contrast (difference in electron density relative to the bulk density). To generate an electron density map of the solute, we assume that the solute is composed of isolated atoms (not chemical bonded), and the deformation electron density is negligibly small. (It should be note that there are lots of work done to incorporate the anisotropicity and asphericity of electron density in order to minimize the deformation density, for example see [219] and references therein.) The density is thus computed by summing up the contribution of all the atoms. In practice, since electron is mostly found near the nucleus, a simple cut-off to decide whether an atom contributes electron density is sufficient. One way that is widely used in X-ray crystallography to compute the electron density around an atom is via the analytic Fourier transform of the atomic scattering factor. [220, 221] Conventionally, the atomic scattering factor was fitted with the sum of some Gaussian terms (with a and b are tabulated constants for each atom, we use here the sum of six Gaussians by Su and Coppens [139]):

$$f\left(\frac{q}{4\pi}\right) = \sum a_i e^{-b_i q^2/16\pi^2} \tag{4.23}$$

Fourier transform of the above formula gives the electron density at a distance r (Å) from the nucleus:

$$\rho\left(r\right) = \sum a_i \left(\frac{4\pi}{b_i}\right)^{3/2} \exp\left(-\frac{4\pi^2 r^2}{b_i}\right) \tag{4.24}$$

The PDDF can also be calculated for the cross-term, $\tilde{F}_{hyd}\left(q\right)\tilde{F}_{solu}\left(q\right)$ or $\tilde{F}_{counterion}\left(q\right)\tilde{F}_{solu}\left(q\right)$. This time the density must come from both solute and water (or counterions):

$$p\left(r\right) = \left\langle \int_V \rho_i\left(\mathbf{r}'\right)\rho_j\left(\mathbf{r}+\mathbf{r}'\right)d\mathbf{r}' \right\rangle \tag{4.25}$$

4.3 Results and Discussions

4.3.1 Validation of the new decomposition scheme

As discussed above, the condition for decomposition of the total scattering intensity into partial intensities is that all the phases should be very close to each other for every orientation of the $\boldsymbol{q}$ vector. To study the behavior of the phases, we must rely on the calculated profiles. We choose to use the calculated SAXS profiles computed by 3D-RISM as they have been shown in the previous Chapter to match the experimental curves up to wide angle region. The phase for each component is computed from the complex amplitude of the corresponding term. The phases are different at each orientation of $\boldsymbol{q}$ (with the same magnitude q), and the average phase is shown in Figure 4.1 for lysozyme and a 25-bp DNA in 100mM NaCl solution. It is expected to see those phases start to deviate from 0 as q increases, but at small angles most of the phases are identical, except for the co-ion Cl^{-} in the DNA case, which is not surprising because Cl^{-} is mostly expelled from the DNA and its contribution is expected to be negative. The total amplitude phase closely tracks the solute phase in the range of $q < 0.5$ Å^{-1} considered here. To make sure that the phases are aligned in all directions, and not just on average, we consider the difference in phase between two amplitudes, α. From Eq. 4.15 one has:

$$\cos\alpha = \frac{F_{solv}\left(\mathbf{q}\right) F_{solu}\left(\mathbf{q}\right)}{\left|F_{solv}\left(\mathbf{q}\right)\right| \left|F_{solu}\left(\mathbf{q}\right)\right|} \tag{4.26}$$

where the solvent amplitude is $F_{solv}\left(\mathbf{q}\right) = F_{hyd}\left(\mathbf{q}\right) + F_{ion}\left(\mathbf{q}\right)$. Figure plots the average value of $\cos\alpha$ over all orientations of the $\boldsymbol{q}$ vector at low angles. The condition of $\cos\alpha \approx 1$ is valid for $q < 0.1$ Å^{-1}. Comparing curves between lysozyme and DNA indicates that the shape at moderate angles ($0.1 < q < 0.5$ Å^{-1}) is system dependent, and for some system the errors will be smaller than for others; in this case, the average error for the DNA is smaller than lysozyme. However, at small angle the errors are vanishingly small, thus validating the decomposition scheme.

Fig. 4.3 illustrates the overall accuracy of Eq. 4.14, comparing the total amplitude $\tilde{F}\left(q\right)$ and the sum of partial amplitudes $\tilde{F}_k\left(q\right)$ for lysozyme and the DNA. The partial amplitudes are defined and calculated as in Section 4.2.2. For Lys, k is *solu* and *hyd*; while for DNA there are also counterion (Na^{+}) and co-ion (Cl^{-}). One can see that the sum of partial amplitudes is a good approximation of the square root of the measured intensity $\tilde{F}(q)$ at least at small angle region ($q < 0.2$ Å^{-1}).

To check whether the $\tilde{F}_{hyd}(q)$ extracted above reflects the distribution of water in the reciprocal space, we compare it with $\sqrt{I_{hyd}\left(q\right)}$ computed directly from the (calculated) water distribution around the solute. The I_{hyd} is calculated by performing the SAXS calculation as usual, but ignoring the solute term (the solute excess form factor is set to 0, *i.e.* it does not interact with the X-ray beam). The result is plotted in Figure 4.4 (left) for lysozyme as the test case. It can be seen that $\tilde{F}_{hyd}\left(q\right)$ from the new analysis scheme is essentially identical with the one computed from the 3-dimensional distribution of water, at least at the small

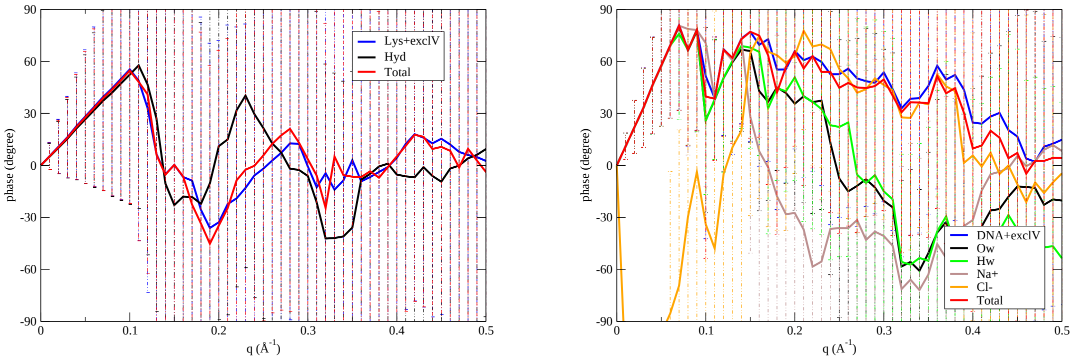


Figure 4.1: Phases of component amplitudes in lysozyme (left) and DNA (right) from the calculated scattering profiles. Each value is the averaged phase all over every $\boldsymbol{q}$ vector, and the error bars report the standard deviation of the distribution. The phase of the total amplitude is also depicted there (red solid line). For both systems, at small angle most of the amplitudes are aligned (*i.e.* in phase), except for the co-ion Cl^{-} in the DNA case, which is excluded from the solute and thus is out of phase. Interestingly, the phases of the total amplitude follow the phase of the solutes very closely.

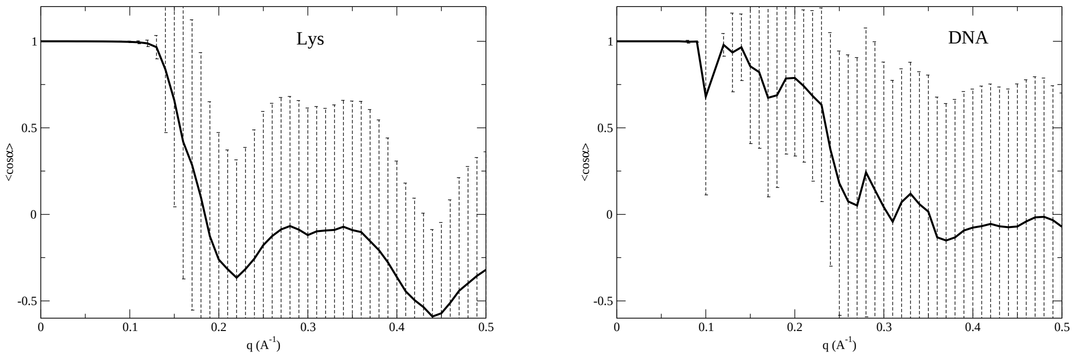


Figure 4.2: Average values of $\cos\alpha$ versus q for lysozyme (left) and DNA (right), where α is the phase difference between $F_{solu}(\mathbf{q})$ and $F_{solv}(\mathbf{q})$ (the standard deviations are also depicted as error bars.) At small angle up to 0.1 Å^{-1} the two amplitudes are nearly in the same phase at every orientation of $\boldsymbol{q}$ with very small error bars, thus validating the decomposition of total intensity at small q region.

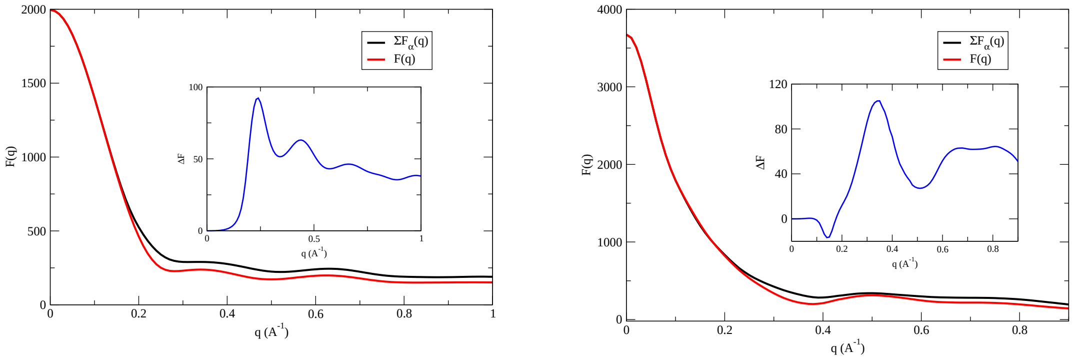


Figure 4.3: Comparison between the "real" excess amplitude $\tilde{F}(q) = \sqrt{I(q)}$ (red) with the sum of component amplitudes $\tilde{F}_k(q) = \sqrt{I_k(q)}$ (black) for lysozyme (left) and DNA (right) using calculated profiles from 3D-RISM. For Lys, k is *solu* and *hyd*; while for DNA there are also counter-ion (Na^+) and co-ion (Cl^-). The difference between these two are plotted in the insets (blue).

angle region. Given SAXS data contain only a few independent values, [208] the fact that we could not capture the whole curve is not discouraging. It is possible that this small angle region is enough to reconstruct the low-resolution PDDF features in real space. We test the quality of $\tilde{F}_{hyd}(q)$ by calculating the PDDF in the real space (details are in Section 4.2.3) and compare this directly with the 3-dimensional distribution. As can be seen in Figure 4.4 (right), the PDDFs from all three approaches are very close to each other. Therefore one can expect that the new analysis $\tilde{F}_{hyd}(q)$ contains enough information to reconstruct the main features of the water distribution.

In the following sections, we apply this new analysis to both calculated and experimental data for proteins and DNA. The goal is to study the quality of the predicted ion and water distributions from different theoretical models, to understand their weaknesses, and to inform future attempts to improve the models and force fields.

4.3.2 Protein test cases

Fig. 4.5 plots the "square-root" subtraction $\tilde{F}_{hyd}$ from SAXS data and those from 3D-RISM and MD for lysozyme and myoglobin (calculated from Eq. 4.16.) There are two features that can be extracted from those curves. The first is the total excess number of water molecules

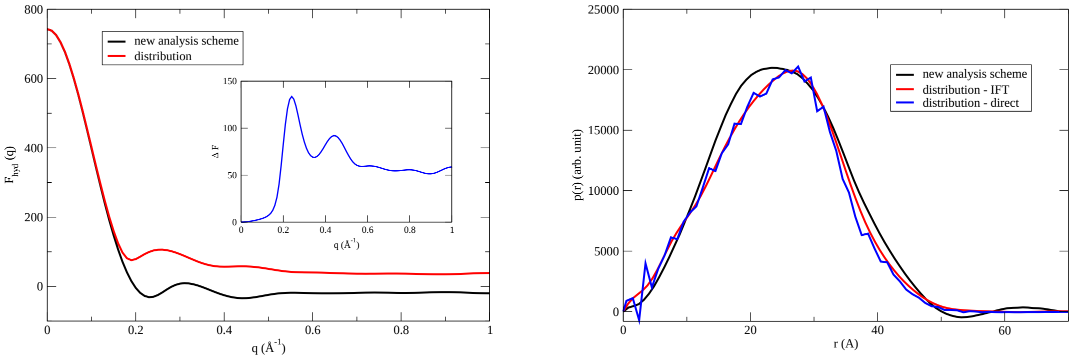


Figure 4.4: (Left) Scattering amplitude of water $\tilde{F}_{hyd}\left(q\right)$ around lysozyme extracted from the square-root subtraction (*black*) in comparison with those computed directly from the water distribution (*red*). The difference between those two, as shown in the inset, is negligible at small angle. (Right) Pair distance distribution function (PDDF) of water in the hydration shells computed for two curves in the left by indirect Fourier transformation of $\tilde{F}^2_{hyd}\left(q\right)$. Also shown in blue is the PDDF calculated by using the direct method in real space (see section 4.2.3 for more details).

in the hydration shell, visible at $q = 0$. Since each water molecule has 10 electrons, $\tilde{F}_{hyd}\left(0\right)$ should be equal to 10 times the number of excess hydration waters N_{hyd}. Second, the shape of the curve contains information about the water distribution in the real space. If $\tilde{F}_{hyd}\left(q\right)$ decays rapidly towards zero, that means the hydration shell is thick. On the other hand, if the curve slowly approaches zero, the hydration water shells are more compact. Of course, it should be easier to explore the latter feature in the real space rather than the reciprocal space using a restricted Fourier transform (IFT) technique.

All RISM closures tend to overestimate $\tilde{F}_{hyd}\left(q\right)$, especially in the small angle region. The higher the order of PSE-n closures, the more serious the overestimation is. This indicates that the water attraction to the solute in RISM is too strong. For example in Lys, RISM-KH overestimates $F_{hyd}(0)$ by about 200, corresponding to ~20 water molecules (Table 4.1). The MD results are in much better agreement with the experiment. There are also some discrepancies around q = 0.2 Å^{-1}, where the new analysis scheme may break down (see Figure 4.2). It is worth noting that although the computed total SAXS profiles of those two proteins from RISM and MD are nearly indistinguishable (see Figure 3.4), the extracted $\tilde{F}_{hyd}$ here is clearly able to separate MD from RISM results. This demonstrates the power

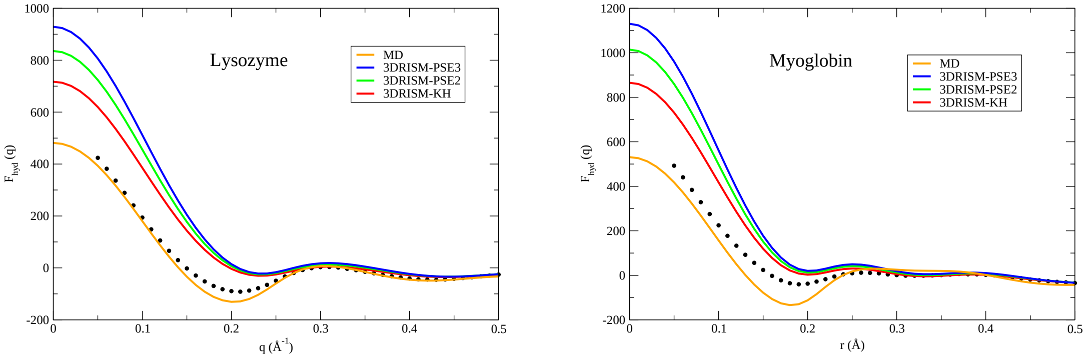


Figure 4.5: $\tilde{F}_{hyd}\left(q\right)$ for lysozyme (left) and myoglobin (right) from SAXS data (black circles, experimental data from Ref. [11]), compared to RISM and MD calculations. The calculation was done as in Eq. 4.16.

Protein	RISM-KH	RISM-PSE2	RISM-PSE3	MD	SAXS
Lysozyme	71.7	83.5	92.9	48.1	50 ± 1
Myoglobin	86.5	101.4	113.0	53.1	60 ± 2

Table 4.1: Number of excess hydration water for lysozyme and myoglobin computed as $N_{ex-hyd} = \tilde{F}_{hyd}\left(0\right)/10$. Those numbers are very close to the values computed by integration all over the hydration shells. For lysozyme, RISM-KH values are 71.7 vs. 71.9.

of the proposed analysis and illustrates its potential use in testing new closures in 3D-RISM.

To get information about the placement of those excess water molecules in real space, we can compute the cross-term $\tilde{F}_{hyd}\left(q\right)\tilde{F}_{solu}\left(q\right)$ (via Eq. 4.18) and transform it to real space to obtain a pair distribution function by using the IFT technique discussed in Section 4.2.3. The results for two proteins, lysozyme and myoglobin, are given in Figure 4.6. Computed SAXS profiles for MD and 3D-RISM are analyzed in the same fashion as described above for the experimental SAXS profiles. The PDDF plot is essentially a distance histogram of hydration water and the solute, weighted by the excess electron density (relative to the bulk solution). Unlike the regular PDDF used in SAXS (which is the Fourier transform of the positive total intensity and should be positive everywhere), the cross-term PDDF here could be negative since $\tilde{F}_{hyd}\left(q\right)\tilde{F}_{solu}\left(q\right)$ is negative at some points. 3D-RISM, regardless of closure, overestimates the hydration water interaction with proteins, whereas results from MD simulation are generally much better. At very small distances, the PDDF contains valuable information about the distances between the solute atoms and neighboring water molecules.

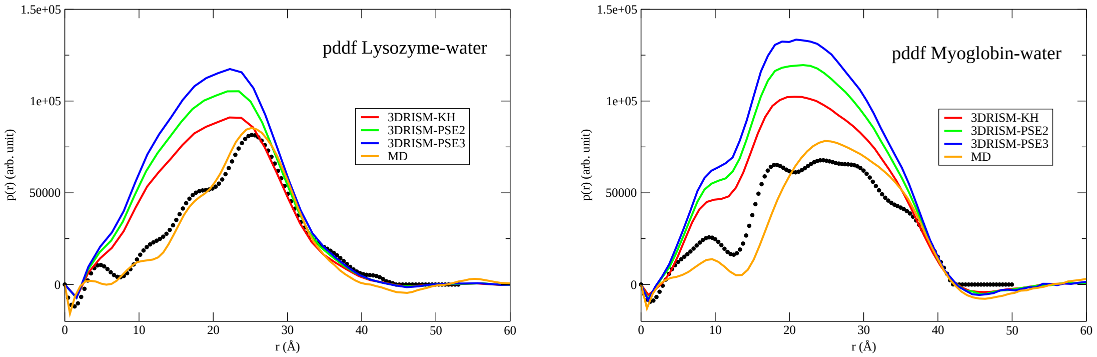


Figure 4.6: Pair distance distribution function (PDDF) of water-solute for lysozyme (left) and myoglobin (right). The PDDFs are computed by IFT-ing the cross-term $\tilde{F}_{solu}(q)\,\tilde{F}_{hyd}(q)$ for experimental data (black, taken from Ref. [11]), MD (orange) and 3D-RISM.

Noteworthy at $r \approx 1-2$ Å, the solute-water PDDF is negative, giving experimental evidence for the so-called "thermal volume", which is the void volume created around the solute due to the mutual vibrations of the solute and solvent as well as structural, packing and steric effect. [222–224] The PDDFs for the two proteins show rich structural features, especially at small distances; this is reminiscent of similar structure seen by Kofinger and Hummer [225] (although their PDDF arises from a Fourier transform of the total intensity, not of a component as we use here). Since only excess waters that are close to the proteins contribute to the PDDF, for nearly spherical proteins, the peak location correlates well with the protein radius.

To make sure the features observed in the PDDF are real, and not artifacts of the IFT technique, we compute directly the PDDF in the real space and compare with the IFT PDDF. Two grids of excess electron density are generated separately for the biomolecule and hydration shells. The PDDF is consequently calculated as a distance histogram between the two grids, weighted by the excess electron density. (See Section 4.2.3 for more details about how we construct the excess density map.) The comparison between the "direct" and IFT PDDF for two proteins using 3D-RISM is shown in Fig. 4.7. Although there is some slight discrepancy at 30–40 Å distance (the IFT moderately underestimates the PDDF in this range), the two methods agree quantitatively, suggesting that the new analysis proposed

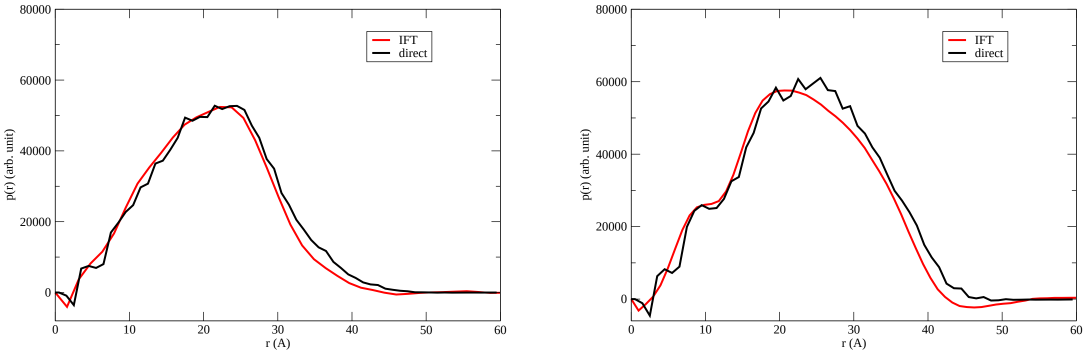


Figure 4.7: Comparison between water-solute PDDF computed by IFT and "direct" method for lysozyme (left) and myoglobin (right). For the "direct" method, two grids of excess electron density are created separately for the solute and hydration shells. The PDDF is then computed by making a distance histogram between the solute and hydration water, weighted with the excess electron density. See Section 4.2.3 for a detailed description of the "direct" method.

here is capable of obtaining a usefully accurate pair distribution.

Generally, we observe that using $q \leq 0.1$ Å^{-1} is enough to construct a "coarse" PDDF for the solute-hydration term (see Fig. 4.8). Using less data leads to the difficulty of converging the IFT procedure, whereas using more data introduces more fine features into the calculated PDDF.

4.3.3 Duplex DNA in salt solutions

If a solute is highly charged, the situation gets more complicated because of the presence of the ionic atmosphere. As we will show below, the contributions of counterions and co-ions can be the same order of magnitude as that of the more numerous hydration water molecules, so that one must account for them in the decomposition. ASAXS is one approach to probe the spatial distribution of ions around DNA. [191, 210] The ASAXS experiment probes the same sample at two different energies, which causes the "effective" number of electron in the interested ion to vary. However, the ASAXS profile does not entirely come from the ion of interest but also has contributions from hydration water-ion cross terms (and solute-ion terms), making it difficult to interpret and draw fruitful conclusions. (One advantage of ASAXS is that the same ion is used.) Another technique is to use heavy ion replacement

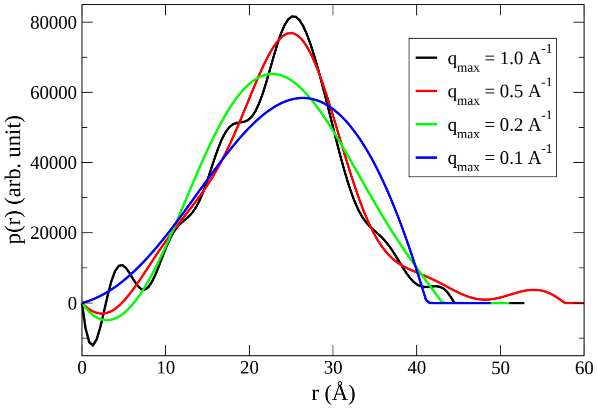


Figure 4.8: Influence of truncation in Fourier space onto the PDDF in real space of lysome-water term $F_{solu}(q)\,F_{hyd}(q)$. The transform is carried out by IFT using experimentally extracted data.

where instead one varies the ion identity to change the contrast. [13] Using a novel analysis technique (instead of the simple subtraction between the two scattering curves as is done in conventional ASAXS), Meisberger *et al.* were able to separate the ion-DNA term from the water-DNA term, and thus could gain insight into the nature of ion cloud around duplex DNA. The method assumes that both the ion and water distributions around DNA are not sensitive to ion type and it has applied successfully to alkali chlorides. (See Section 4.3.4 below for the relationship between our analysis and the method from Meisburger *et al.*) It is not clear how accurate this assumption is or whether the same assumption would be valid for highly charged ions (such as Mg^{2+}, Sr^{2+} ...) since the interaction between those ions with nucleic acids are expected to be ion-dependent. [226–230]

Here, we apply our analysis method to a 25 base-pair duplex DNA. Figure 4.9 shows the ion-solute cross-terms for experimental and calculated SAXS data of the duplex DNA in 100mM RbCl or 10mM $SrCl_2$. As for hydration waters, we extract the total number of excess ions and a qualitative description of their distribution in real space. (Probing the ion cloud around charged biomolecules in a very dilute solution (as the 10mM $SrCl_2$ solution) using MD simulation is prohibitively expensive, [9, 99, 100, 195] so we only report 3D-RISM results for Sr^{2+}.) As reported earlier, [9] and shown here in Figure 4.9 (at $q = 0$) and Table

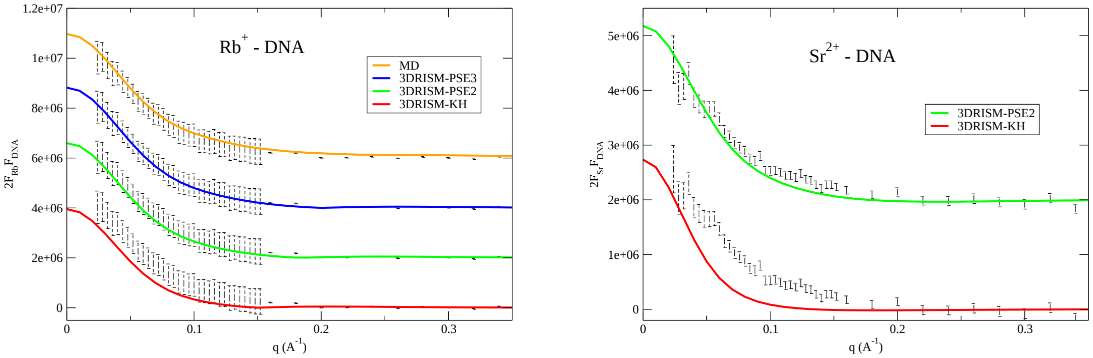


Figure 4.9: Rb^{+}–DNA (left) and Sr^{2+}–DNA (right) cross-terms from the calculated (color solid lines) and experimental (black dots) ASAXS data of DNA in 100mM RbCl or 10mM SrCl_2 solutions; experimental data is from Ref. [12]. The curves were offset to facilitate visual comparison. From those curves, one could extract the excess number of ions (at $q = 0$) and qualitatively infer about the ion cloud around the DNA.

4.2, 3D-RISM (with high order closures) and MD simulation (at least for monovalent ions) are able to reproduce accurately the excess number of ions around DNA, including both monovalent and divalent ions.

We also performed IFT to obtain the PDDFs of the ion-DNA crossterm, which are shown in Figure 4.10. The curves for Rb^{+} look encouraging, especially for the MD simulation and for high order closures in RISM. High order closures tend to place more ions closer the DNA, which is more consistent with MD and experiment. The discrepancies at large distances (especially for Sr^{2+}) probably rise from the lack of data at very small angles, hindered by the geometry of the beam stop in the SAXS experiment setup. The errors at large r of the PDDF curves could also arise from the way theoretical models approximate ion-solute interaction. The PDDF curves for divalent ions show large deviations from the experiment despite having somewhat reasonable agreement in the Fourier space from PSE2 closure. This highlights the fact that the number of excess ions should not be used solely to characterize the ion cloud. Instead, information about the shape of the ion cloud should be also taken into account. The 3D-RISM model, in its current form, is known to have difficulties with divalent ions, [121, 231] perhaps resulting from the lack of polarization effects. [232, 233] More work currently is underway to test new ion models in RISM calculations.

System		N_{cation}
DNA/100mM RbCl	ASAXS	37 ± 2
	MD	37.95
	RISM-PSE3	36.77
	RISM-PSE2	35.08
	RISM-KH	30.16
DNA/10mM $SrCl_2$	ASAXS	20 ± 2
	RISM-PSE2	22.24
	RISM-KH	19.14

Table 4.2: Number of excess ions around a -48 charged duplex DNA in 100mM RbCl or 10mM $SrCl_2$ solutions, computed as $N_+ = \tilde{F}_{ion}(0)/Z^+$ from the new analysis scheme. Those numbers are very close to the values computed by integration over all space. For example, values computed at RISM-PSE3 for Rb^+ are 36.77 (from the $q = 0$ limit) vs. 36.85 (from directly integrating the distribution).

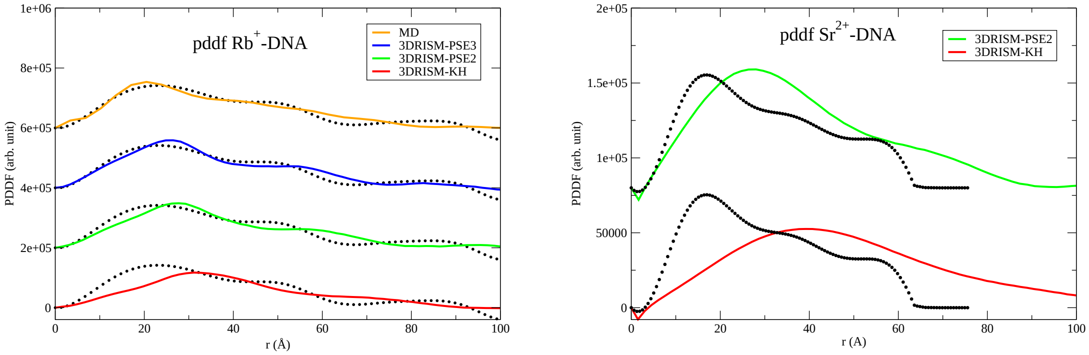


Figure 4.10: PDDF of Rb^+–DNA (left) and Sr^{2+}–DNA (right) obtained by inverse Fourier transformation of the ion-DNA cross-terms $\tilde{F}_{ion}(q)\,\tilde{F}_{DNA}(q)$. Results are shown for experimental data (black, taken from Ref. [12]), MD (orange) and 3D-RISM. The curves are offset to facilitate visual comparison. No MD data is reported for $SrCl_2$ due to the high cost of simulation for dilute (10 mM) solution.

To determine the water hydration term, the co-ion term (in this case Cl^-) needs to be taken into account. We note that the contribution of the solvent to the total intensity is based on the scattering contrast (relative to the bulk concentration), and since the co-ions are entirely excluded from such a highly charged DNA, the contribution of this deficit to the X-ray scattering difference is therefore significant. A simple calculation shows that with the DNA considered here (which has a -48e charge), $N_{Rb} = 37$ leads to $N_{Cl} = -11$ and therefore $\tilde{F}_{Rb}(q = 0) = 1332$ vs. $\tilde{F}_{Cl}(q = 0) = -198$. The co-ion term will contribute even more strongly if lighter counter-ions (Na^+, K^+ ...) and/or heavier co-ions (Br^-, I^- ...) are used. (A recent ion counting experiment emphasizes the importance of co-ion identity to the ion atmosphere around nucleic acids. [211]) Here, we estimate the co-ion contribution using the simple model described in Section 4.2.2.2.

The water term $\tilde{F}_{hyd}$ is then determined and plotted in Figure 4.11 for DNA in two different salt solutions. The PDDFs of water-DNA in two different salt solutions are computed by performing IFT of $\tilde{F}_{hyd}\tilde{F}_{DNA}$ and are shown in Figure 4.12. All RISM and MD results overestimate the number of excess water compared with experiment, however at different level for different salts. The numbers of waters in the two case remain relatively unchanged in theoretical predictions, whereas it varies a lot with experimental data: the experimental values for N_{wat} are extrapolated from those curves to $q = 0$ and are ~ 70 and 110 for RbCl and $SrCl_2$, respectively, *i.e.* the difference is around 1.6 water molecules per base pair. This difference could potentially come from the fact that fewer Sr^{2+} are required to neutralize the DNA than Rb^+ (see Table 4.2), leading to fewer ions accumulating near the DNA surface and therefore providing more space for water. Also, Sr^{2+} is expected to have denser and stronger hydration shells than Rb^+, which will be dragged along the ions towards the DNA. The much smaller concentration of $SrCl_2$ compared with RbCl (10mM vs. 100mM) is probably another factor leading to fewer ions accumulating near the DNA surface.

4.3.4 Comparison to an alternate decomposition approach

Instead of using different beam line energies to change the efficient contrast of the interested ion, one may instead vary the ion itself and assume both the ion and water distributions are

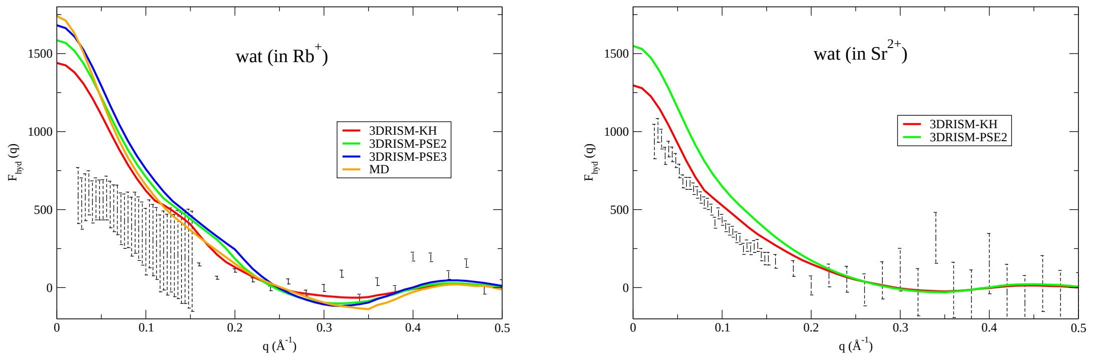


Figure 4.11: Water hydration term $\tilde{F}_{hyd}$ extracted from SAXS data of the 25-bp DNA in 100mM RbCl (left) or 10mM SrCl$_2$ (right) (computed as described in Eq. 4.21) for experimental SAXS (black error bars taken from Ref. [12]), MD (orange) and 3D-RISM.

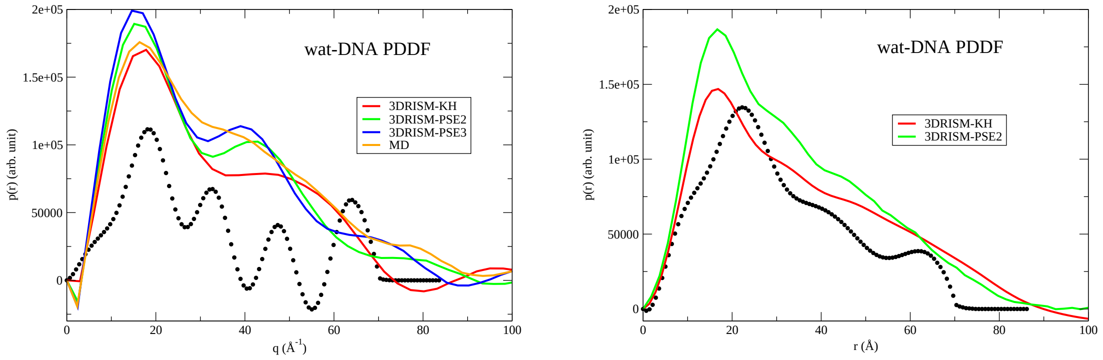


Figure 4.12: PDDF of water-DNA in RbCl 100mM (left) or SrCl$_2$ 10mM (right), calculated by performing IFT for $\tilde{F}_{DNA}(q)\,\tilde{F}_{hyd}(q)$. Black circles are from the decomposition of experimental data reported in Ref. [12].

unchanged. Meisburger *et al.* express the total SAXS intensity as: [13]

$$I\left(q\right)=\delta_{solu}^{2}P_{solu}\left(q\right)+2\delta_{solu}\left(\delta_{ion}N_{ion}\right)P_{solu-ion}\left(q\right)+\left(\delta_{ion}N_{ion}\right)^{2}P_{ion}\left(q\right)\tag{4.27}$$

where δ is the scattering contrast and P is the partial scattering form factor (which has a range from 0 to 1). For the solute, $\delta_{solu}=Z_{solu}+N_{wat}Z_{wat}$, therefore it includes all the contribution from water coming from both from the excluded volume and hydration shells. The first term in Eq. 4.27 is equivalent to the sum of the solute and water in our approach:

$$\begin{aligned}\tilde{F}_{solu+hyd}^{2}\left(q\right)&=\left(\tilde{F}_{solu}\left(q\right)+\tilde{F}_{hyd}\left(q\right)\right)^{2}\\&=\left[\left(Z_{solu}+N_{excl}Z_{wat}\right)\tilde{f}_{solu}\left(q\right)+N_{hyd}Z_{wat}\tilde{f}_{hyd}\left(q\right)\right]^{2}\end{aligned}\tag{4.28}$$

with $\tilde{f}$ is the normalized $\tilde{F}$ ($\tilde{f}\left(0\right)=1$) and the number of excess waters N_{wat} is partitioned into excluded volume contribution N_{excl} and hydration contribution N_{hyd}. If one assumes $\tilde{f}_{solu}\left(q\right)=\tilde{f}_{hyd}\left(q\right)$ then the right hand side of Eq. 4.28 could re rewritten as $\left(Z_{solu}+N_{wat}Z_{wat}\right)^{2}\tilde{f}_{solu}^{2}\left(q\right)$, leading to

$$P_{solu}\left(q\right)=\tilde{f}_{solu}^{2}\left(q\right)\tag{4.29}$$

The third term in Eq. 4.27 is equivalent to our ion term:

$$\tilde{F}_{ion}^{2}\left(q\right)=\left(N_{ion}Z_{ion}\tilde{f}_{ion}\left(q\right)\right)^{2}\tag{4.30}$$

Compare Eq. 4.30 with the last term of Eq. 4.27 leads to

$$P_{ion}\left(q\right)=\tilde{f}_{ion}^{2}\left(q\right)\tag{4.31}$$

The cross-term in eq. 4.27 is then

$$P_{solu-ion}\left(q\right)=\tilde{f}_{solu}\left(q\right)\tilde{f}_{ion}\left(q\right)\tag{4.32}$$

Eqs. 4.29, 4.31 and 4.32 relate the two different approaches to extract ion distribution from X-ray scattering experiment. It is obvious that $P\left(q\right)$ and $\tilde{f}\left(q\right)$, which are both defined to be within 0 and 1, are basically the same entity. $P\left(q\right)$ couples with the intensity I, while $\tilde{f}\left(q\right)$ couples with the partial amplitude $\tilde{F}$. The advantage of using $\tilde{f}\left(q\right)$ instead of $P\left(q\right)$ is that every component is separated completely from each other and there is no cross-term; therefore one only needs n $\tilde{f}\left(q\right)$ instead of $\sim n^{2}$ $P\left(q\right)$ to specify a system with

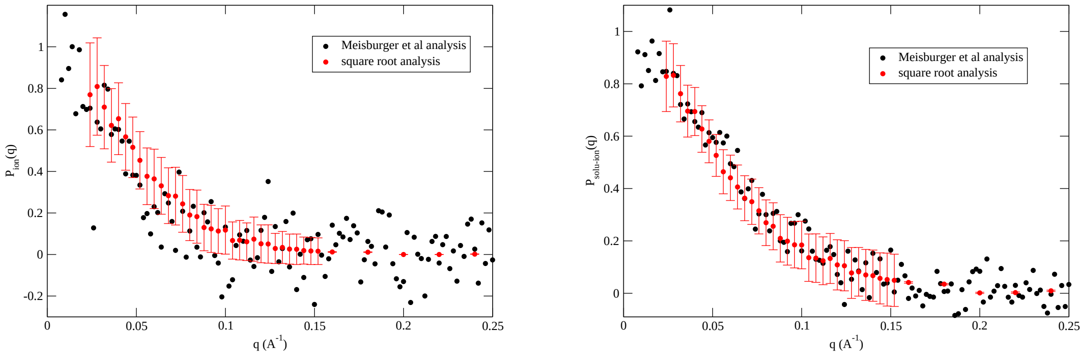


Figure 4.13: $P_{ion}(q)$ (left) and $P_{DNA-ion}(q)$ (right) comparison between two analysis schemes for the DNA SAXS data: black circles are from Meisburger *et al.* [13], red circles are from the method proposed here. Error bars are computed by propagating the errors from experimental intensity data. $P_{DNA-ion}(q)$ is actually a normalized $F_{ion}(q)\,F_{DNA}(q)$ curve (see Figure 4.9), so the error bars here are the same error bars in Figure 4.9 after normalization. They were computed by propagating the error bars when (square root) subtracting the two experimental intensities I_{off} and I_{on} for $\tilde{F}_{ion}$, times with $\tilde{F}_{DNA}$.

n components. In addition, the partial amplitudes are additive while the intensities are not, which is potentially easier to work with.

Figure 4.13 compares $P_{ion}(q)$ and $P_{solu-ion}(q)$ from Meisburger *et al.* and our analysis. $P(q)$ from our method is directly related to $\tilde{f}(q)$ which is nothing but the normalized $\tilde{F}(q)$. It is apparent that the two methods agree quantitatively despite the fact the Meisburger *et al.* implicitly assume $\tilde{f}_{solu}(q) = \tilde{f}_{hyd}(q)$, which seems to be reasonable (the hydration shell shape should be somewhat similar to the solute shape).

4.4 Conclusions

Water molecules and ions around biomolecules often play a crucial role in function. Here we propose a new analysis scheme for X-ray scattering data to extract information about how water molecules and ions distribute around the solute. We show that although the analysis requires some approximation, it is accurate enough to obtain reliable partial scattering intensities in Fourier space as well as distribution functions in real space. The resulting distributions could then be used to study the dynamic nature of the solvation shells, for instance via time-resolved scattering techniques. [234–236] It could also be used to test the accuracy of theoretical predictions, eventually to make improvements in how those theories treat water molecules, ions and cosolvents in general. Comparing theory vs. experiment for individual interaction terms (as in Figs. 4.6, 4.10 and 4.12) is likely to be more helpful in assessing the strengths and weaknesses of theoretical models than just making comparisons to the complete SAXS profile. The proposed analysis complements recent experimental techniques (such as ion counting [7] and anomalous SAXS [106]) by providing not only the number excess of particles but also their distribution in real space. It is, however, worth restating the fact that our decomposition requires an independent knowledge of the structure of solute, which is assumed to be rigid; it cannot be used (in its current form) for systems with significant conformational heterogeneity or disorder.

To illustrate the new analysis, we extract the water hydration distribution around two proteins - lysozyme and myoglobin - from regular X-ray scattering profiles. Comparison between those experimental distributions (extracted from SAXS data) and the calculated distributions from 3D-RISM and MD simulation reveals that MD simulation accurately accounts for water in terms of both number of excess water and its real space distribution, whereas 3D-RISM overestimates the number of excess waters leading to the accumulation of water hydration near the proteins. This overestimation could come from the way RISM treats water molecules and the approximation made in the theory involving the bridge function. First, the water models used in RISM are modified versions of SPC/E and TIP3P, which include Lennard–Jones parameters for hydrogen atoms, making the water-water interaction much weaker and inadvertently increasing the water-solute interaction. Second, the RISM closure equations are approximate in nature since the bridge function ideally should be an infinite series of functionals representing 3-body and higher order interactions. Simply ignoring the bridge function (as in HNC-like closures used here) probably perturbs the overall interaction of the whole system.

For highly charged systems such as the DNA duplex, both MD simulation and 3D-RISM (with high order closures) are capable of capturing the ion cloud of counterions around the DNA, again both in terms of number excess and the real space distribution. Water molecules, on the other hand, are predicted to be too strongly attracted towards the DNA, presumably the phosphate groups, by both MD simulation and 3D-RISM. This is an unexpected result and highlights the need to recalibrate nucleic acid force fields. There are also some studies in the literature reporting a misbalance between solute-water and solute-solute interaction, which is probably relevant to the situation here. [55, 59, 62, 237, 238]

We have chosen in Eq. 3.37 to included the “excluded volume” effect (the fact that solvent is excluded from the interior of the biomolecule) into our definition of $F_{solu}(\mathbf{q})$. This was an arbitrary choice, but driven by the fact that the “excluded volume” effect can be easily computed from a known structure, and by the perspective that the “interesting” parts of hydration are those that take place outside the solute interior. But other choices, such as including only the first term of Eq. 3.37 in $F_{solu}(\mathbf{q})$ are possible, and do not change the analysis method here in any fundamental way. Future studies should help to determine the relative strengths and weaknesses of different decomposition methods.

Chapter 5

Partial molar volume and number of hydration water

5.1 Introduction

It is widely accepted that solute-solvent interaction is extremely important in governing the biomolecule shape and function. Recent force field developments have attempted to incorporate solute-water interaction in one form or another in the parametrization process. [4, 238–241] As we show in chapter 4, the distribution of water molecules and ions around macromolecules can be extracted from experimental SAXS measurements. Those distributions alongside with the number of excess water of hydration can serve as additional data to calibrate the solute-solvent interaction. Since absolutely calibrated SAXS data are only available for a limited number of systems, it is valuable if other experimental sources could also provide the same information.

In this chapter, we discuss the use of the so-called Kirkwood–Buff (KB) theory to obtain similar information about solute-solvent interaction from thermodynamic measurements. The partial molar volume (PMV) is shown to contain exactly the same amount of information as in SAXS data (at $q = 0$) (namely the number of hydration water and ions around the solute), and is available for most systems due to its ease of measurement. Using the solute PMV as an additional data for force field derivation, we illustrate that generally the solute-solvent interaction becomes more balanced and the number of excess water around the solute is in better agreement with experimental data. One advantage of PMV and the number of hydration water is that they are relatively easy to measure in the experiment and are available for any system, while hydration free energy, for instance, is only available for small molecules and ions, and mostly not accessible (or very hard to measure) for proteins and nucleic acids.

We start this chapter by briefly introducing the KB theory for liquid state with the focus on calculating PMV from the pair distribution function $g(r)$. We then show that current force fields reasonably reproduce the experimental PMV for lowly charged proteins, although there is still room for development. Calculated PMV of charged residues and molecular ions, however, are in clear disagreement with the experiment. We then develop a new set of parameters for molecular ions following the IPolQ protocol [242] and show that the predicted PMV for those ions are greatly improved. The parameters are then used to adjust the phosphate groups in the DNA duplex while leaving all the backbones, sugars and bases unchanged. With this refined phosphate group, the calculated PMV and number of excess hydration water of a duplex DNA are shown to approach the experimental values from SAXS data.

5.2 Kirkwood–Buff theory

5.2.1 Overview

Theories of liquid state and of liquid mixtures could be classified into two main themes. One direction attempts to relate the interaction between molecules to the structure of liquid (such as the integral equation theory discussed in chapter 2). The other approach is the Kirkwood–Buff (KB) theory which establishes the structure of liquid to thermodynamic parameters without the need to know the interaction within the system. The KB theory provides a direct relationship between thermodynamic properties (such as isothermal compressibility, partial molar volumes, derivatives of chemical potentials) and the so-called KB integrals, defined as:

$$G_{ij} = 4\pi \int_0^\infty \left[g_{ij}(r) - 1\right] r^2 dr \tag{5.1}$$

where g_{ij} is the pair correlation function between two species i and j in the open system (grand canonical ensemble with T, V, μ–the chemical potential–being kept constant). Since $g(r)$ is readily obtained from RISM theory and MD simulations, KB theory is thus an excellent tool to link the “structure” calculated in RISM and MD simulation to thermodynamic measurements. It should be noted that KB theory is considered an exact theory, *i.e.* its derivation entirely comes from mathematical manipulations of statistical mechanics relationships without making any assumptions. Disagreements between predicted and experimental values of thermodynamic quantity are thus rooted from the approximations one uses to obtain the correlation function, and not from the KB theory itself.

To establish the relationship between thermodynamic quantities and the pair correlation function, KB theory uses the fluctuation in the number of the particles as an intermediate. Starting with the definition of $g\left(r\right)$ in an open system, one has:

$$G_{ij} = V\left(\frac{\left\langle N_iN_j\right\rangle - \left\langle N_i\right\rangle\left\langle N_j\right\rangle}{\left\langle N_i\right\rangle\left\langle N_j\right\rangle} - \frac{\delta_{ij}}{\left\langle N_i\right\rangle}\right) \tag{5.2}$$

where $\left\langle N\right\rangle$ is the number fluctuation in a fixed volume V, i and j denote the species $(i, j = 1, 2, ..., n)$ and δ is the delta function. Next, considering the grand canonical partition function as:

$$\Xi\left(T, V, \mu_1, ..., \mu_n\right) = \sum_n Q\left(T, V, N_1, ..., N_n\right)\exp\left(\beta\sum_k \mu_k N_k\right) \tag{5.3}$$

The average number $\left\langle N_i\right\rangle$ in the system is:

$$\begin{aligned}\left\langle N_i\right\rangle &= \Xi^{-1}\sum_n N_iQ\left(T, V, N_1, ..., N_n\right)\exp\left(\beta\sum_n \mu_k N_k\right)\\ &= k_BT\left[\frac{\partial\ln\Xi\left(T, V, \mu_1, ..., \mu_n\right)}{\partial\mu_i}\right]_{T,V,\mu_{j\neq i}}\end{aligned} \tag{5.4}$$

Differentiating Eq. 5.4 with respective to μ_j we have:

$$\begin{aligned}k_BT\left(\frac{\partial\left\langle N_i\right\rangle}{\partial\mu_j}\right) &= \Xi^{-1}\sum_n N_iN_jQ\left(T, V, N_1, ..., N_n\right)\exp\left(\beta\sum_n \mu_k N_k\right) - \left\langle N_i\right\rangle\left\langle N_j\right\rangle\\ &= \left\langle N_iN_j\right\rangle - \left\langle N_i\right\rangle\left\langle N_j\right\rangle\end{aligned} \tag{5.5}$$

Combining Eqs. 5.2 and 5.5 gives:

$$\begin{aligned}B_{ij} &\equiv \frac{k_BT}{V}\left(\frac{\partial\left\langle N_i\right\rangle}{\partial\mu_j}\right)_{T,V,\mu_{k\neq j}}\\ &= \rho_i\rho_jG_{ij} + \rho_i\delta_{ij}\end{aligned} \tag{5.6}$$

Eq. 5.6 is then used to obtain other thermodynamic observables such as the isothermal compressibility χ_T, partial molar volumes, derivatives of chemical potentials:

$$\chi_T = \frac{\det B}{k_B T \sum_{i,j} \rho_i \rho_j B_{ij}} \tag{5.7}$$

$$\bar{V_\alpha} = \frac{\sum_i \rho_i B_{i\alpha}}{\sum_{i,j} \rho_i \rho_j B_{ij}} \tag{5.8}$$

$$\begin{aligned} \mu_{\alpha\beta} &\equiv \left(\frac{\partial \mu_\alpha}{\partial N_\beta}\right)_{T,P,N_{\gamma\neq\beta}} \\ &= \frac{k_B T}{V \det B} \frac{\sum_{i,j} \rho_i \rho_j \left(B_{\alpha\beta} B_{ij} - B_{i\alpha} B_{j\beta}\right)}{\sum_{i,j} \rho_i \rho_j B_{ij}} \end{aligned} \tag{5.9}$$

with B_{ij} is the element of the matrix B, as defined in Eq. 5.6. Eqs. 5.7–5.9 are the main results of KB theory. It is important to note that the pair correlation functions in KB theory do not depend explicitly on the orientation of the particles, but the theory is valid for any kind of particles, not necessarily spherical particles. There is no assumption about the interaction between the particles, whether it is pairwise or not. Additionally, KB theory is also valid for quantum systems although here we only discuss classical systems.

5.2.2 Partial molar volume

The partial molar volume is defined at constant temperature and pressure as following:

$$\begin{aligned} \overline{V}_i &= \left(\frac{\partial V}{\partial n_i}\right)_{T,P,n_{j\neq i}} \\ &= \left(\frac{\partial \mu_i}{\partial P}\right)_{T,n} \end{aligned} \tag{5.10}$$

with V is the volume of the system, and the notion of $n_{j\neq i}$ indicates that every number of molecules in the system except the protein are held constant. PMV can be computed directly from the KB integrals as in Eq. 5.8. In this chapter, we are interested in the PMV of the biomolecule at the infinite dilution. For two-component systems, it is convenient to define two quantities η and ς as:

$$\eta = \rho_A + \rho_B + \rho_A \rho_B \left(G_{AA} + G_{BB} - 2G_{AB}\right) \tag{5.11}$$

$$\varsigma = 1 + \rho_A G_{AA} + \rho_B G_{BB} + \rho_A \rho_B \left(G_{AA} G_{BB} - G_{AB}^2\right) \tag{5.12}$$

so that the thermodynamic quantities in Eqs. 5.7 and 5.8 can be rewritten as:

$$\chi_T = \frac{\varsigma}{k_B T \eta} \tag{5.13}$$

$$\bar{V}_i = \frac{1 + \rho_j \left(G_{jj} - G_{ij}\right)}{\eta} \tag{5.14}$$

In the limit $\rho_B \to 0$ (B is the biomolecule of interest), Eqs. 5.13 and 5.14 become:

$$\begin{aligned} \lim_{\rho_B \to 0} \chi_T &\equiv \chi_T^0 \\ &= \frac{1 + \rho_A G_{AA}}{k_B T \rho_A} \end{aligned} \tag{5.15}$$

$$\begin{aligned} \lim_{\rho_B \to 0} \bar{V}_A &\equiv \bar{V}_A^0 \\ &= \frac{1}{\rho_A} \end{aligned} \tag{5.16}$$

$$\begin{aligned} \lim_{\rho_B \to 0} \bar{V}_B &\equiv \bar{V}_B^0 \\ &= \frac{1 + \rho_A \left(G_{AA} - G_{AB}\right)}{\rho_A} \end{aligned} \tag{5.17}$$

Combining Eqs. 5.15–5.17 gives the formula for the partial molar volume of the solute at the infinite dilution:

$$\bar{V}_B^0 = k_B T \chi_T^0 - \bar{V}_A^0 \rho_A G_{AB} \tag{5.18}$$

Eq. 5.18 could be generalized in the solution of n components:

$$\begin{aligned} \bar{V}^0 &= k_B T \chi_T^0 - \sum_i \bar{V}_i^0 \rho_i G_i \\ &= k_B T \chi_T^0 - \sum_i \bar{V}_i^0 N_i \end{aligned} \tag{5.19}$$

where $N_i = \rho_i G_i = \rho_i \int \left[g_i\left(\mathbf{r}\right) - 1\right] d\mathbf{r}$ is the number of excess solvent i around the solute, χ_T^0 is the isothermal compressibility of the solvent (without the solute presence) and $\bar{V}_i$ is the partial molar volume of solvent i (again, without the solute). The first term in Eq. 5.19 arises due to the translational degree of freedom of the solute. [222, 223, 243] This term contributes to the solute PMV even if the solute collapses into a single point with no dimension and no interaction with the surrounding solvent, *i.e.* the term is equal to the volume increase of the system once an inert point is introduced. Due to its small magnitude (around 1 Å^3) compared with the biomolecules considered here (with the volume is on the order of thousands Å^3), it is safely ignored.

The PMV of a biomolecule is directly related to the numbers of excess solvent around it and thus contains valuable information about the solute-solvent interaction (although in an integral way). In the next section we explore a possibility of using PMV to recalibrate solute-solvent interactions.

5.3 Methods and computational details

PMV computation

The PMV can be computed as in Eq. 5.19 (ignore the ideal term which is very small compared to the volume of biomolecules):

$$\begin{aligned} \bar{V} &= -\sum_i \bar{V}_i^0 \rho_i G_i \\ &= -\sum \bar{V}_i^0 N_i \end{aligned} \tag{5.20}$$

Given $g(\mathbf{r})$ is the direct output, it is straightforward to compute PMV from 3D-RISM. The PMV can also be equivalently computed from the direct correlation function $c(\mathbf{r})$ as: [244,245]

$$\bar{V} = -k_B T \chi_T^0 \sum_i \rho_i \hat{c}_i(0) \tag{5.21}$$

where $\hat{c}_i(0) = \int c_i(\mathbf{r})\, d\mathbf{r}$.

To compute PMV in MD simulation, one could construct a 3-dimensional $g(\mathbf{r})$ maps from the simulation trajectory and then use Eq. 5.20. This approach, however, requires much more computational expense if there are cosolvents (with small concentrations) present in the system and does not incorporate the solute flexibility into the calculation. Another method uses the definition of PMV (as in Eq. 5.10) as the volume increase when the solute is present. [246] This method computes the volume difference between a simulation of the solute + n_w waters and a simulation of n_w water molecules:

$$\bar{V} = \langle V_{n+1} \rangle - \langle V_n \rangle \tag{5.22}$$

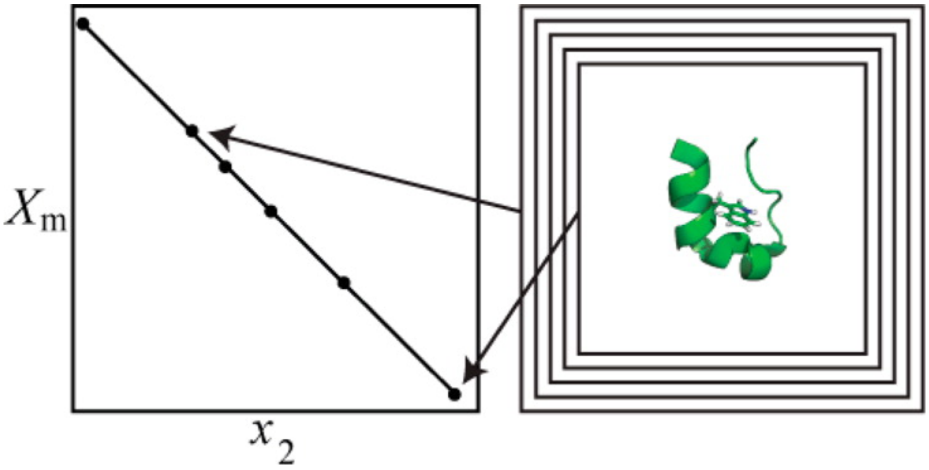


Figure 5.1: Illustration of the partial molar volume calculation, reprinted from Ploetz and Smith. [14]

where $\langle\rangle$ indicates isothermal–isobaric (NPT) ensemble average. However, this method is shown to suffer from large statistical errors because the value of interested is evaluated by the difference between two large numbers. [247] Other methods that are much more expensive such as free energy perturbation exist. [248]

We use here the method proposed by Ploetz and Smith to compute PMV from MD simulation in addition to the KB method. [14] The effective concentration of the solute is altered by varying the number of solvent molecules used in the simulation, while keeping only one solute molecule in the simulation boxes. The solute PMV is then calculated by linear-fitting the volumes in those simulations (see illustration in Figure 5.1). If the molar volume is defined as $V_m = V/N$ where V is the volume of the simulation box and N is the total number of molecules in the simulation (regardless they are solute or solvent) then it can be written as:

$$V_m = V_m^* + a x_{solu} \tag{5.23}$$

where V_m^* is the molar property of the pure solvent and x_{solu} is the mole fraction of the solute. Plotting V_m vs. x_{solu} will give both the slope a and V_m^*. The PMV of the solute is then determined as:

$$\bar{V} = a + V_m^* \tag{5.24}$$

Method		$\bar{V}$ (10^3 Å^3)
Kirkwood–Buff		17.47
Ploetz–Smith	Rigid	17.32
	Flexible	16.91

Table 5.1: Partial molar volume of lysozyme computed by two different methods using MD simulation: Kirkwood–Buff approach via the correlation function $g(\mathbf{r})$ and the finite difference approach by Ploetz and Smith. With the finite difference approach, two separate calculations are performed in which the solute is either fixed or fully flexible. All simulations are performed with the AMBER ff14SB force field and SPC/E water.

5.4 Results and Discussions

5.4.1 Protein PMV

First we compare two different methods used in MD simulation to compute PMV: i) KB approach as in Eq. 5.20 and ii) the finite difference method proposed by Ploetz and Smith. [14] We also test the effect of solute flexibility on the PMV calculation for a relatively rigid protein. Results are shown in Table 5.1 for lysozyme. (Note that all PMVs reported in this work are per molecule, not per mole.) It is obvious that the two approaches give almost identical PMV for the fixed protein. Introducing the solute flexibility into the calculation slightly reduces the computed PMV by only 2%. Given that the finite difference method requires multiple simulations while the KB approach only needs a single run and the solute flexibility is of minor importance for rigid proteins, we feel confident to use the KB approach for further study.

5.4.1.1 Protein test cases

Table 5.2 and Figure 5.2 present the PMV for various proteins computed by 3D-RISM and MD with the KB approach. Most proteins are relatively rigid so that the solute flexibility does not contribute much to the PMV. The results indicate that both 3D-RISM and MD simulation are capable to reproduce the experimental PMV well. There is a trend of MD simulation slightly overestimates the PMV while 3D-RISM slightly underestimates it, especially for large proteins. Such a small errors indicate a subtle misbalance in solute-solvent interaction and will be discussed in more details in next sections, where one partitions the PMV into hydration shell and excluded volume contributions.

Protein	PDB ID	3DRISM-KH	MD	Exp
Pti	4PTI	7.58	8.06	7.78
α-Lactalbumin	1ALC	16.5	17.3	17.1
Ribonuclease A	3RN3	15.9	16.2	16.0
Lysozyme	6LYZ	17.2	17.5	17.4
Adenylate kinase	3ADK	26.0	27.0	26.7
Papain	1PPN	27.9	29.2	28.0
Concanavalin A (monomer)	2CNA	30.1	32.7	31.1
Elastase	3EST	31.3	32.8	31.4
Carbonic anhydrase B	2CAB	33.7	35.2	34.9
Subtilisin	2SBT	33.0	35.5	33.4
Rhodanese	1RHD	39.6	41.5	40.6
Carboxypeptidase A	2CTB	41.6	43.1	42.0

Table 5.2: Comparison between partial molar volumes of proteins (10^3 Å^3) computed by 3DRISM-KH and MD using the KB approach with the experiment (taken from Ref. [1]). Data for lysozyme calculated from SAXS.

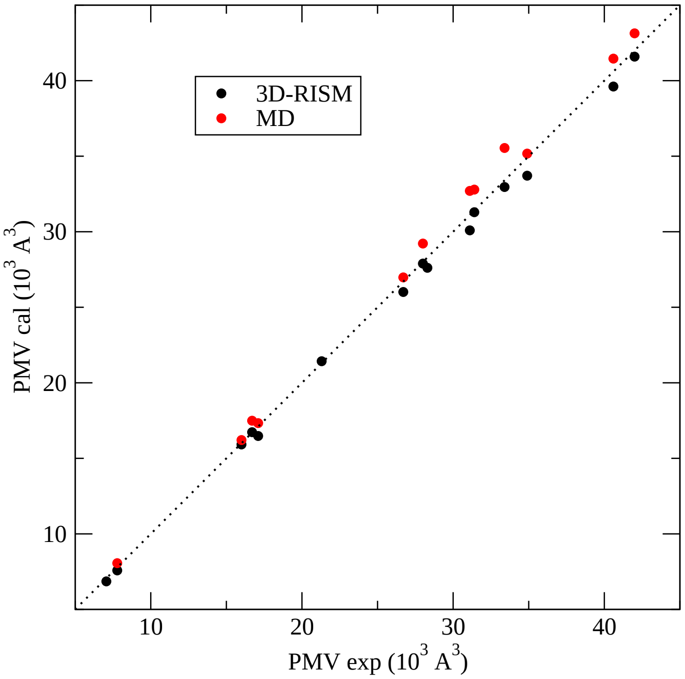


Figure 5.2: Comparison of partial molar volumes computed from 3D-RISM and MD simulation (using the AMBER ff14SB force field) with experimental data, taken from Ref. [1]. The proteins are listed in Table 5.2.

5.4.1.2 Protein force field and water models

If there is only water in the solvent then the PMV is related directly to the number of excess water as:

$$\bar{V} = -\bar{V}_{wat} N_{wat} \tag{5.25}$$

where $\bar{V}_{wat} \approx 30$ Å^3 is the PMV of pure water. N_{wat} is always negative for "large" molecules because when put in the solution, they displace some water molecules, and therefore increasing the apparent volume of the system. N_{wat} can be further divided into contributions from the excluded volume and hydration water (as done in Section 4.2). The excluded volume contribution is the number of bulk water displaced by the "geometric" (or "intrinsic") volume of the protein, given the surrounding water molecules are not affected by the protein presence. This contribution only depends on the topology of the solute and is "easily" (though arbitrarily) computed once the solute structure is known, thus it is of minor interest here. On the other hand, the hydration contribution N_{hyd} results from the protein-water interaction and thus will be the main focus in this chapter.

Table 5.3 reports the PMV of lysozyme computed by KB approach using different protein force fields and water models. The number of excess water is separated into excluded volume and hydration contributions. Here we compute the excluded volume using a grid-based method as done by Voss and Gerstein with a probing radius of 1.4 Å (see Figure 5.3). [16] We also present the PMV from SAXS data (inferred from the excess number of water). The results show that the predicted PMVs (and N_{wat}) generally agree with the values obtained by SAXS and are insensitive to force fields and water models used. However, more careful inspection shows that the dominant contribution to the PMV arises from the non-interesting excluded volume, while the more "interesting" hydration contribution is very small. Comparison between calculated and SAXS N_{hyd} reveals that there are subtle differences between force fields and water models. For example, a change of water models leads to 10-20% variation of N_{hyd} computed by MD simulation. Such a difference cannot be observed in PMV. On the other hand, the water model does not affect RISM PMVs, all three calculations give very similar values. Interestingly, CHARMM-36 underestimates N_{hyd} while OPLS/AA overestimates it. Again, such errors are hard to detect if one compares N_{wat} and

Force field	Water model	N_{wat}	N_{hyd}	$\bar{V}$ (10^3 Å^3)
AMBER ff14SB [249]	SPC/E	-583.5	27.6	17.47
	TIP3P	-580.6	30.5	17.38
	TIP4P/EW	-577.1	34.0	17.29
AMBER ff14ipq [241]	TIP4P/EW	-586.7	24.4	17.57
CHARMM-36 [250]	TIP3P	-607.8	3.3	18.20
GROMOS 53A6 [251]	SPC/E	-587.2	23.9	17.58
GROMOS 54A7 [252]	SPC/E	-586.6	24.5	17.56
KB FF [239, 253]	SPC/E	-580.7	30.4	17.39
OPLS/AA [254]	TIP4P	-559.8	51.3	16.76
AMOEBA 09 [255]	AMOEBA	-575.0	36.1	17.22
AMOEBA 13 [256]	AMOEBA	-578.3	32.8	17.31
3D-RISM-ff14SB	cSPC/E	-574.9	36.2	17.21
	cTIP3P	-574.5	36.6	17.20
	cOPC3	-575.5	35.6	17.23
SAXS		-581 ± 1	30 ± 1	17.40 ± 0.03

Table 5.3: Numbers of excess water N_{wat}, numbers of excess water in the hydration shells N_{hyd} and partial molar volume of lysozyme calculated by MD simulation and 3D-RISM using different force fields and water models.

PMV instead.

5.4.1.3 Electrostatic vs. non-electrostatic

Since the number of excess waters in the hydration shell contains information about the solute-solvent interaction, we attempt to study the effect of electrostatic interaction between solute-solvent on N_{hyd} by the following: a similar calculation is repeated to compute N_{hyd} but this time all the partial atomic charges on the solute atoms are set to zero. The number of excess hydration water this time ($N_{non-elec}$) only comes from the van der Waals interaction between the protein and solvent and the fact that the protein is present in the solution. The difference between N_{hyd} and $N_{non-elec}$ is solely caused by the electrostatic interaction and will be called N_{elec}. This approach has been used by Hirata and coworkers to study the hydration shell of proteins. [245]

Figure 5.4 shows the partition of N_{hyd} for various proteins of different sizes computed by 3D-RISM. It is interesting that even in the absence of electrostatic interaction between proteins and water molecules, there is still an excess number of hydration water accumulating around the hypothetically neutral proteins. The presence of a neutral solute introduces

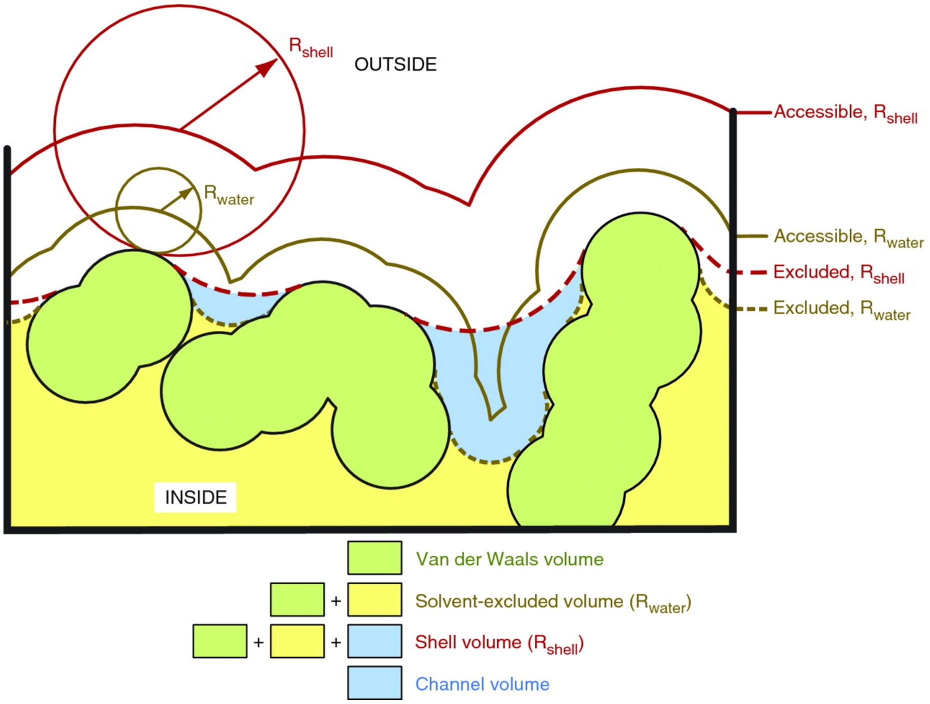


Figure 5.3: Illustration of the excluded volume based on Richards' rolling probe definition. [15] Reprinted from Voss and Gerstein, [16] by permission of Oxford University Press.

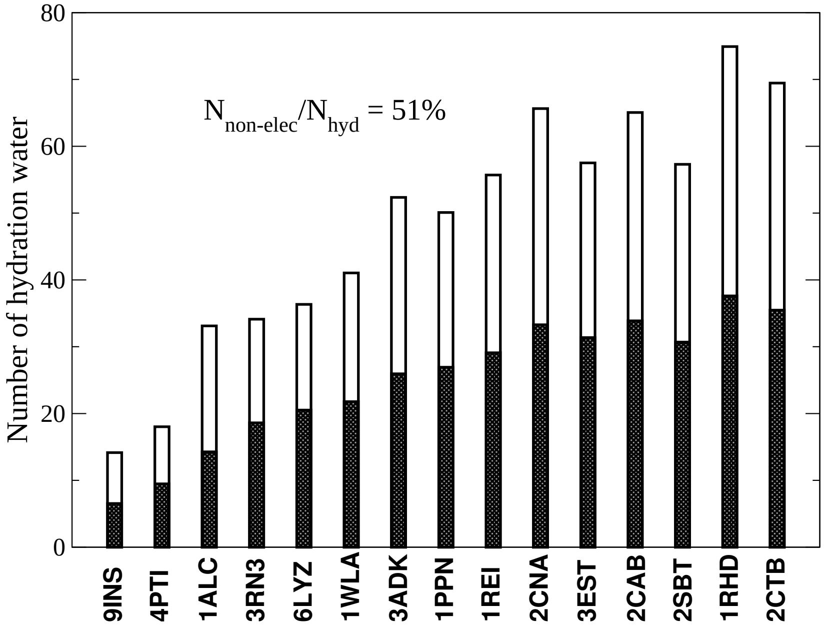


Figure 5.4: Partition of N_{hyd} into electrostatic (blank) and non-electrostatic (filled) contributions for proteins. Results are computed with 3D-RISM coupled with AMBER ff14SB force field and cSPC/E water model.

around half of the excess number of water, and turning on the electrostatic interaction between solute and solvent contributes another half. It should be noted that aside from the van der Waals interaction, the fact that the solute is present in the solution also contributes to $N_{non-elec}$. Considering a monatomic liquid with inert atoms such as Ne or Ar. If one integrates the (excess) radial distribution function starting not from zero but from the atom radius ($\int_{r_a}^{\infty} 4\pi \left[g\left(r\right) - 1\right] r^2 dr$), then one always obtains a positive number. The same situation also happens in the proteins here since we perform the integration outside the atomic radii of the solute atoms. The results here are qualitatively in agreement with those from Merzel and Smith. Using MD simulation, they found that around two-thirds of the water density increase around proteins are merely caused by the protein surface existence. [39] This contribution arises even if the surrounding water molecules were not perturbed by the protein presence.

5.4.2 Small molecules and ions

Given the number of hydration water is hindered by a huge excluded volume term as in proteins, we next turn to small molecules and ions to minimize the impact of the arbitrary definition of the protein surface. Also, any change or recalibration of water-solute interaction should start at small molecules.

5.4.2.1 Amino acid side-chain and backbone analogs

The amino-acid side-chain and backbone analogs are the classical models for force field development since they are the building blocks of proteins. We here try to compare the performance of two different Amber force fields in terms of reproducing the PMV of those analogs: ff14SB and ff14ipq. [241, 249] It should be emphasized that the parametrization protocols for those two force fields are very different. While ff14SB is the more up-to-date version of the “regular” AMBER force fields (only adjusting the side chain and backbone dihedral parameters to better describe the protein conformations) and thus inherit most of the solute-solvent interaction from the old force fields, ff14ipq is a totally different route. The charges on the solute are fitted to reproduce the solvent reaction field in explicit solvent in a

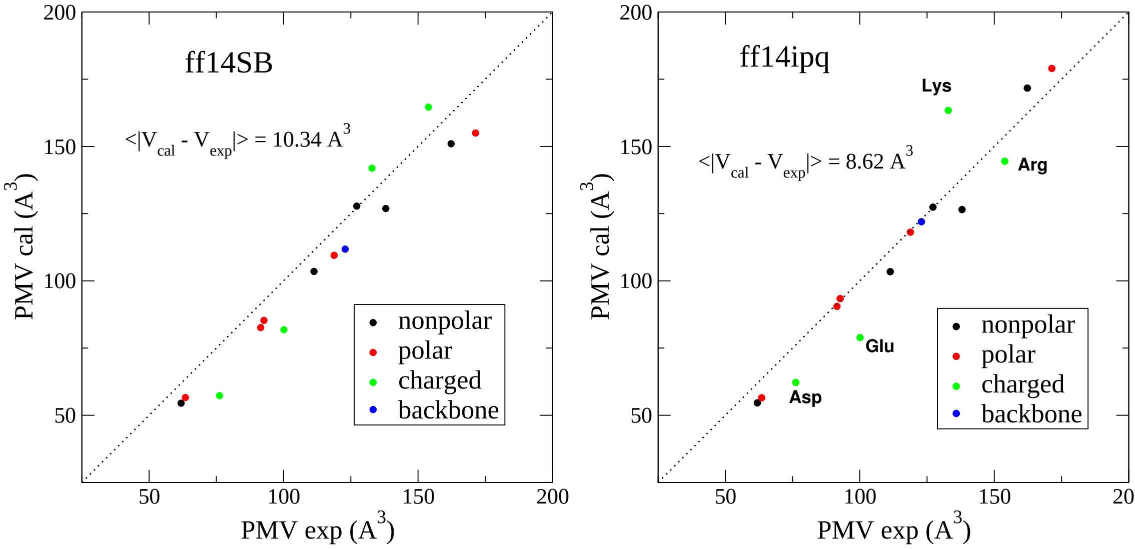


Figure 5.5: Partial molar volumes of amino-acid side-chain and backbone analogs computed from 3D-RISM using two different AMBER force fields: ff14SB and ff14ipq. Experimental data are from Refs. [17–19].

self-consistent manner. [242] The Lennard–Jones parameters are then adjusted to calibrate the solute-solvent interaction by bringing the solute hydration free energy to match with experiment. The ff14ipq force field is, thus, the first Amber force field that calibrates the solute-solvent interaction in the parametrization steps. Note that this trend becomes more popular recently, from fixed-charge to polarizable force fields. [238, 257]

Figure 5.5 compares the PMV of the amino-acid side-chain and backbone analogs for ff14SB and ff14ipq with experiment. Despite of not being parametrized to match experimental PMV or hydration free energy, ff14SB reproduces very well PMVs for the protein building blocks, although it slightly underestimates the experimental data, meaning that the analog-water interaction is modeled somewhat stronger than it should be. The ff14ipq, as expected, replicates most of the PMVs from experiment except for charge residues. It is important to stress that even a slight improvement of PMV can be essential to model the solute-water interaction since the PMV value is hindered by a large excluded volume contribution (as discussed in Section 5.4.1.2).

Ion	Optimized ΔG_{hyd}	Target ΔG_{hyd}	Exp. ΔG_{hyd}
AcO^-	-90.71	-89.8	-77.7[a]
ClO_3^-	-72.41	-71.7	-59.5[b]
ClO_4^-	-65.31	-53.7	-41.6[b]
$HCOO^-$	-90.43	-88.4	-76.3[a]
OH^-	-115.99	-116.9	-104.8[a]
SH^-	-85.18	-84.3	-72.2[a]
SCN^-	-72.42	-71.7	59.5[b]
NO_3^-	-77.48	-76.4	-64.3[b]
HCO_3^-	-90.44	-84.8	-72.7[b]
HSO_3^-	-75.50	-60.7	-48.5[b]
$H_2PO_4^-$	-111.38	-115.9	-103.7[b]
NH_4^+	-71.56	-73.1	-85.2[a]
Gdm^+	-53.97	-55.2	-67.3[c]
$(CH_3)_4N^+$	-41.67	-41.1	-53.2[b]
CO_3^{2-}	-333.48	-331.2	-306.9[b]
SO_4^{2-}	-273.94	-275.0	-250.7[b]
HPO_4^{2-}	-327.51	-323	-299[d]
PO_4^{3-}	-670.21	-689.9	-653.4[b]

Table 5.4: Hydration free energy ΔG_{hyd} (kcal/mol) of ions. Target ΔG_{hyd} is the target value for the computed HFEs to match. The target values are obtained from experimental values after subtracting the contribution of the interfacial potential jump (see more details in Appendix 6). Optimized ΔG_{hyd} is the final HFE after the IPolQ protocol. Experimental data taken from Ref. a [2], Ref. b [3], Ref. c [4], Ref. d [5]. See Appendix 6 for more details about the experimental ion HFEs and corrections made during the computation of HFE.

5.4.2.2 Molecular ions

As discussed in Section 4.3.3, the interaction between DNA-water is modeled too strong in current force fields. Given the solute-solvent interaction can be adjusted by the IPolQ protocol, we perform the calculations for molecular ions with the aim to expand the protocol for ions and more importantly to seek a way to recalibrate the DNA-water interaction. The chosen ions are both biologically relevant (acetate, phosphate, guanidinium) and inorganic in nature and are listed in Table 5.4. The IPolQ procedure is carried out for every ion to obtain the converged partial atomic charges. The Lennard–Jones parameters are then adjusted so that the calculated ion hydration free energies match with the experiment. The details of the IPolQ protocol and free energy calculation can be found in Appendix 6.

Table 5.4 reports the hydration free energies after IPolQ protocol for all ions. In each ion, only a single key atom is chosen to adjust the LJ σ parameter (for instance, oxygen

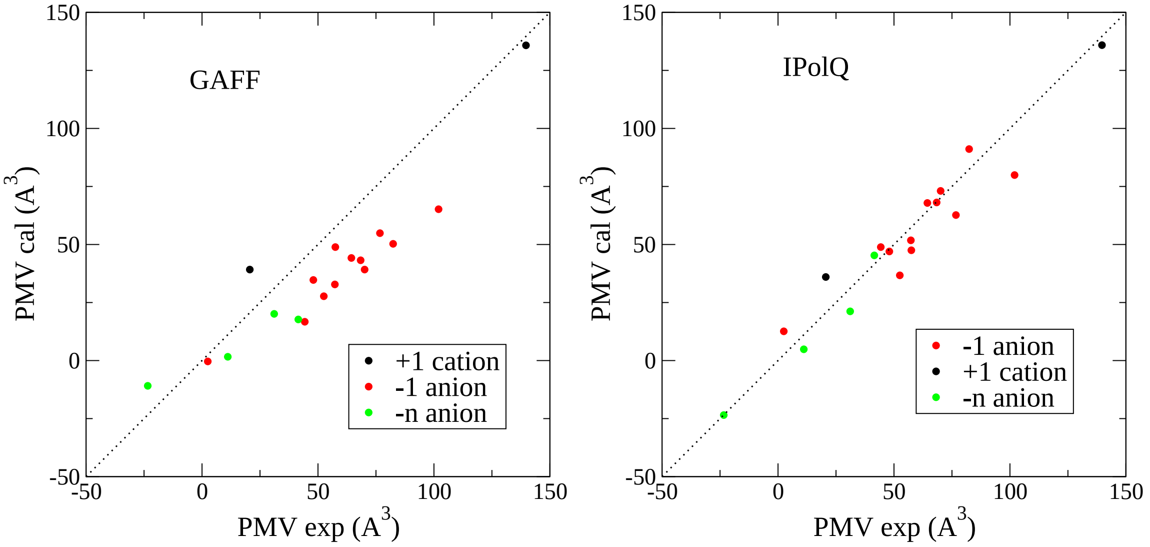


Figure 5.6: Partial molar volumes of molecular ions computed from 3D-RISM using GAFF 1.5 force field and the refined IPolQ-charge model. All experimental data are from Ref. [3].

in carboxylate and phosphate). We constrain ourselves to adjust only σ, while leaving ϵ unchanged. Generally, a good match between IPolQ and experimental HFEs can be obtained if we allow σ to vary within 10-15%, including "important" ions such as acetate, guanidinium and phosphate (-1). This range (10-15%) is actually used in the current ff14ipq force field. However, not all HFEs can be brought back to experimental values and we decide not to "overfit" them.

The new ion parameters are then subjected to PMV calculations using RISM and compared with experiment (results shown in Figure 5.6). It is clear that the refined IPolQ models give improved PMVs compared with the original GAFF parameters, indicating that the ion-water interaction is much more balanced in the new charge model. Most of the outliers have the HFEs that are hard to match in the IPolQ procedure. The results here indicate a strong correlation between HFE and PMV; and since HFE for large biomolecules are not available (or not easily measurable) one can replace it with PMV as a criteria for force field development.

	N_{Rb^+}	N_{Cl^-}	N_{wat}	N_{hyd}	$\bar{V}$ (10^3 Å^3)
3D-RISM KH	30.16	-17.84	-442.5	73.7	13.0
3D-RISM PSE2	35.08	-12.92	-427.9	88.3	12.1
3D-RISM PSE3	36.77	-11.23	-418.2	98.0	11.6
3D-RISM PSE4	37.72	-10.28	-426.4	89.8	11.8
MD	37.95	-10.05	-440.7	75.46	12.3
SAXS data	39 ± 2	-9 ± 2	-508 ± 11	8 ± 11	14.2 ± 0.5

Table 5.5: Number of excess ions and water around a 25bp duplex DNA in 100mM RbCl solution from 3D-RISM and MD compared with SAXS data. (The values for SAXS obtained from the procedure described in Section 4.2.2.) Also shown the number of hydration water N_{hyd} and partial molar volume of the DNA computed from those excess numbers from Eq. 5.19.

5.4.3 Nucleic acid PMV and hydration

As discussed in Section 4.2.2, the water interaction with the DNA is currently modeled too strong for both 3D-RISM and MD using parm-bsc0 force field. As shown in Table 5.5, the number of excess ions from RISM with high closures and MD are in great agreement with SAXS data (at $q = 0$) but the number of excess water is overestimated ~ 70 molecules (or 1.4 water molecules per base). This leads to an underestimation of PMV and the DNA is "effectively" smaller because it attracts too much water molecules around it, making the system volume smaller. The strong interaction presumably comes from the phosphate groups which have an effective charge of -1. SAXS data of DNA duplex in different salt solutions (Na$^+$, K$^+$, etc) and from different nucleic acid sequences (RNA, hybrid DNA:RNA, etc) also give the same conclusion (data not shown).

Since the PMV of ions can be corrected by an IPolQ parameter optimization in Section 5.4.2.2, we consider to recalibrate the DNA-water interaction by just repeating the IPolQ procedure on a phosphate group analog while leaving all other backbone, sugar and base parameters unchanged. This comes from our assumption that those large errors in the excess water counts are mainly caused by the phosphate groups. We show below that additional optimization steps for backbone, sugar and especially the base parameters are probably needed. We choose dimethylphosphate (DMP) as the phosphate group analog and start the IPolQ protocol as in Section 5.4.2.2. The parameters of the optimized DMP are given in Table 5.6.

Atom	q (e)	σ_{orig} (Å)	σ_{opt} (Å)
P	1.2715	2.1000	2.1000
OP	-0.8697	1.6612	1.9500
OR	-0.4676	1.6837	1.7713
C	-0.0049	1.9080	1.9080
H	0.0688	1.3870	1.3870

Table 5.6: IPolQ optimized parameter for dimethylphosphate. Only the σ of oxygen atoms are optimized. All the ε parameters are kept unchanged.

We then integrate the new phosphate parameter into the 25bp duplex DNA and redo the calculation to compute the number of excess solvent for 3D-RISM. It should be noted that although we have not obtained the MD results for the refined DNA, we expect that the number of excess ions should be around the values of PSE3 closure while the number of excess water molecules should be around the values of KH closure (as in Table 5.5 for the unmodified DNA). The results are shown in Table 5.7. It is interesting that the modified phosphate groups do not change the number of excess ions around the DNA but cause a sharp decrease of the number of excess hydration water N_{wat} and N_{hyd}. Figure 5.7 presents the radial distribution function of water molecules and Rb^+ ions around the phosphate groups. As expected, since we increase the oxygen radii in the phosphate groups, the peaks in the distribution shift slightly towards larger distance for both the water and ion. The fact that there is a strong decrease in the water count and almost no change in the ion count mostly dues to the difference in the density between those two. Note that the ratio of the difference of water count and the difference of ion count is around ~500, similar to the ratio of the two bulk densities of water and ion. Although the number of excess water gets smaller, there is still around 20 water molecules overestimated by 3D-RISM (and possibly MD). This highlights a need to refit the partial charges on other moieties as well, especially the bases.

An interesting feature that makes DNAs are slightly different from proteins is the number of excess hydration waters caused by the electrostatic interaction N_{elec} contributes around two-thirds of the total excess hydration waters (Table 5.7), higher than those in proteins which are about a half (Figure 5.4). For the DNA with modified phosphate groups, we observe that the electrostatic contribution totally dominates the water count, contributing

	Parm-bsc0	IPolQ phosphate	SAXS
N_{Rb^+} (PSE3)	36.85	36.77	39 ± 2
N_{Cl^-} (PSE3)	-11.2	-11.2	-9 ± 2
N_{wat} (KH)	-442.5	-478.5	-508 ± 11
N_{hyd} (KH)	73.7	37.7	8 ± 11
N_{elec} (KH)	49.3	34.4	N/A
$\bar{V}$ (10^3 Å^3) (KH)	13.0	14.1	14.2 ± 0.5

Table 5.7: Number of excess ions and water molecules around a 25bp duplex DNA in 100mM from 3D-RISM. Note that the reported values for ions are taken from PSE3 closure, while results for water molecules and PMV are from KH closure. The DNA here has the phosphate groups with the modified IPolQ charge and LJ parameters. Also shown are the number of hydration water N_{hyd} (caused by a full interaction) and N_{elec} (caused by only electrostatic interaction, see Section 5.4.1.3 for more details about this partition).

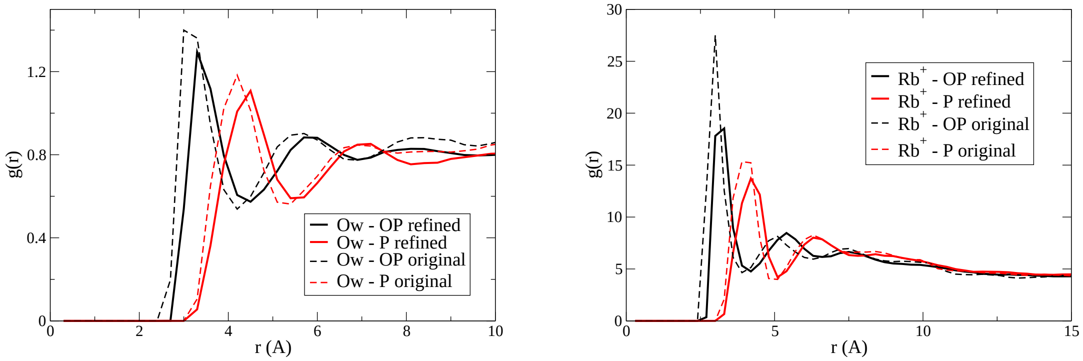


Figure 5.7: Pair distribution function $g(r)$ between water-phosphate (left) and Rb$^+$-phosphate in a 25bp duplex DNA using the original parm-bsc0 (dash lines) and refined phosphate (solid lines) parameters.

more than 90% of the total number of excess hydration water. It is worthy noting that the IPolQ charges are derived to implicitly account for the polarization effect which is probably important in the DNA here. [242] However, more works are required to verify the magnitude of the electrostatic contribution, and to understand more quantitatively about nucleic acid hydration in general.

5.5 Conclusions

We show in this chapter that the Kirkwood–Buff theory can be used to relate the partial molar volume with the number of excess solvent around the solute and thus they are both can be used in force field development to recalibrate the solute-solvent interaction (in addition to the hydration free energy and others). Different methods to calculate the PMV in MD simulation and 3D-RISM are tested and they give essentially similar results. For relatively rigid biomolecules, the solute flexibility contributes negligibly to the calculated PMV and thus one could safely used a "fixed" solute to facilitate numerical computations.

The solute PMV and the number of excess water molecules can be conceptually divided into the excluded volume and hydration shell contributions. For macromolecules, the excluded volume contribution dominates in the PMV. This is somewhat unfortunate because the excluded volume term does not contain useful information about solute-solvent interaction and only depends on the solute topology. The more "interesting" hydration shell term is small and requires extra care to obtain useful information. We illustrate that most current widely used protein force fields provide reasonable estimates of the hydration shell contribution, although there should be more works to improve the solute-solvent interaction balance. Using the Amber ff14ipq force field that explicitly calibrates the solute-solvent interaction by matching the calculated hydration free energy of the solute with experimental data, the PMVs of amino acid side-chain and backbone analogs are in better agreement with the experiment.

We apply the same procedure as in ff14ipq force field to derive a new set of parameters for molecular ions and show that the computed PMVs for those ions are in great improvement compared with the initial and unbalanced force field. As a proof of concept, we proceed to reoptimize the partial atomic charges and LJ parameters for the phosphate group in the nucleic acids by using dimethylphosphate as the analog. The number of excess ions around the DNA with the optimized parameters is found almost identical while the number of excess hydration waters approaches SAXS data in the right direction. We therefore posit that additional charge optimization should also be carried out for other fragments as well (such as backbones, sugars and especially the bases). It is important to stress that our "optimized" parameters are not useful and cannot be used for dynamics study of DNA because it still requires a refit of bonded parameters as well (bond, angle and dihedral terms) and is therefore extremely costly and requires serious validation. However, given that the nucleic acid structure depends moderately on water interaction strength, we believe that such a misbalance should be corrected in the near future.

Chapter 6

Concluding remarks and future directions

In this dissertation, we study the solvent distribution around biomolecules using a combination of molecular dynamics simulation, integral equation theory and X-ray scattering experiment. The X-ray profiles provide important information into how the biomolecules modify the bulk solvent and thus can be used to benchmark solvation methods.

We describe a method to calculate X-ray scattering intensity from atomic models of proteins and nucleic acids in the solution in chapter 3. We show that although the hydration model in 3D-RISM is far from perfect but it provides estimates of useful accuracy that agree better with experiment for a number of test cases than do the predictions of simple competing models, and rival the results of much more expensive MD simulation. The 3D-RISM is particularly attractive for cases when there are both ions and water in the environment, since there are few existing implicit models that describe both, and equilibration of ion densities in MD simulations can be difficult to achieve. We consider only the "forward" problem of estimating SAXS profiles based on an input structure; the "inverse" problem of constructing a structure or ensemble consistent with a given profile is more challenging, and is generally problem-specific. Our computation is fast enough to allow one to average over many solute configurations, or to use SAXS results (perhaps in combination with other restraints) to construct ensembles of configurations consistent with the data (as done in [198, 199, 258]). The characterization of the solvent perturbation used here relies on a thermally-averaged density profile, and appears to be only appropriate for $q < 1.5$ Å^{-1}. At wider angles, fluctuation in the solvent densities (not just the average density) become important, and a different type of theory is needed. (At high angles, errors in the 3D-RISM description of pure water may also be a factor limiting the application of this model.) Nonetheless, this range of scattering angles covers a large fraction of reported experimental profiles, and our model should be of considerable use.

We then propose in chapter 4 a novel analysis scheme for X-ray scattering to extract information about how water molecules and ions distribute around the biomolecules. We show that although the analysis requires some approximation, it is accurate enough to obtain reliable partial scattering intensities in Fourier space as well as distribution functions in real space. The resulting distributions could then be used to test the accuracy of theoretical predictions, eventually to make improvements in how those theories treating water molecules, ions and cosolvents in general. The proposed analysis complements recent experimental techniques (such as ion counting and ASAXS) by providing not only the number excess of particles but also their distribution in real space. It is, however, worth restating the fact that the current form of our decomposition analysis requires an independent knowledge of the structure of solute, which is assumed to be rigid in this work; it cannot be used (without further modifications) for systems with significant conformational heterogeneity or disorder. The analysis is illustrated by extracting the water and ion distribution around two proteins (lysozyme and myoglobin) and a DNA duplex. Comparison between those experimental distributions (extracted from SAXS data) and the theoretically predicted distributions reveals that for lowly charged proteins, MD simulation accurately accounts for water in terms of both number of excess water and its real space distribution, whereas 3D-RISM overestimates the number of excess waters leading to an accumulation of water hydration near the proteins. For highly charged systems such as the DNA duplex, both MD and 3D-RISM (with high order closures) are capable of capturing the ion cloud of counterions around the DNA (again, both in terms of number excess and the real space distribution). Water molecules, on the other hand, are predicted to be attracted towards the DNA too strongly, presumably the phosphate groups, by both MD and 3D-RISM. This is an unexpected result and highlights a need to recalibrate nucleic acid force fields.

In chapter 5, we show that the number of excess solvent particles around the solute can also be obtained with partial molar volume measurements, in addition to SAXS experiment. Different procedures to calculate the PMV via Kirkwood–Buff theory and MD simulation are tested to show that they give essentially identical results. For relatively rigid biomolecules, the solute flexibility contributes negligibly to the PMV and thus one could safely use a fixed conformation of the solute to reduce the computational expense. We further conceptually divide the PMV and the number of excess solvent particles into contributions from the solute excluded volume and the hydration shell and find that the excluded volume term dominates in the macromolecule PMV. The more "interesting" hydration shell term contains the number of excess hydration water and is thus valuable for force field development. We illustrate that most current protein force fields provide good estimates of the hydration shell term, although there is still room for improvement. For nucleic acids, the number of excess hydration water is overestimated by current force fields. As a proof of concept, we reoptimize the non-bonded parameters for the phosphate groups in the DNA by using dimethylphosphate ion as the analog. The computed number of excess hydration waters around the "optimized" DNA approaches the experimental value, while the number of excess ions remain unchanged. Additional charge optimization for other moieties such as backbones, sugars and bases is therefore probably needed. Our parameters, however, are not useful and cannot be used for dynamic study of nucleic acids because it requires a complete refit of bonded terms as well, and thus is extremely costly and requires serious validation thereafter. However, given that the nucleic acid structure depends tightly on the solute-solvent interaction, we believe that such a misbalance should be corrected in the near future.

Appendix

Pressure calculation in RISM

The pressure can be calculated by 3 different ways: virial route, compressibility route and energy route. [41]

Virial route

The pressure P can be calculated from the radial distribution function $g(r)$ of the solvent via the virial equation (or pressure equation). For monatomic liquid, it can be written as:

$$\frac{\beta P}{\rho} = 1 - \frac{2\pi\beta\rho}{3}\int_0^\infty \frac{\partial u(r)}{\partial r} g(r) r^3 dr \tag{6.1}$$

where $u(r)$ is the interaction potential.

Compressibility route

Definition of the isothermal compressibility

$$\chi_T = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T \tag{6.2}$$

χ_T can be computed from $h(r) = g(r) - 1$ as:

$$\rho k_B T \chi_T = 1 + 4\pi\rho \int h(r) r^2 dr \tag{6.3}$$

Energy route

Relationship between the pressure and the Helmholtz free energy A: [121]

$$\begin{aligned} p &= \frac{G-A}{V} \\ &= -\frac{A}{V} + \sum_i \rho_i \mu_i \end{aligned} \tag{6.4}$$

The Helmholtz free energy can be computed as: [123, 125]

$$\begin{aligned}\frac{A}{V} &= 2\pi \sum_{\alpha}^{N_{site}} \sum_{\gamma}^{N_{site}} \rho_\alpha \rho_\gamma \int_0^\infty \left(\frac{h_{\alpha\gamma}^2}{2} - c_{\alpha\gamma} - \frac{\left(t_{\alpha\gamma}^*\right)^{n+1}}{(n+1)\,!} \Theta\left(t_{\alpha\gamma}^*\right) \right) r^2 dr \\ &\quad + \frac{1}{4\pi^2} \int_0^\infty \left\{ Tr\left(\boldsymbol{\omega}\mathbf{c}\boldsymbol{\rho}\right) + \ln\det\left(\mathbf{I} - \boldsymbol{\omega}\mathbf{c}\boldsymbol{\rho}\right) \right\} k^2 dk\end{aligned} \tag{6.5}$$

Appendix

IPolQ protocol for ion parametrization

The ion structures are built and optimized with Gaussian at B3LYP/6-311G**. Atomic charges and LJ parameters are first taken directly from GAFF 1.5. The IPolQ procedure is then carried out for each ion, following Cerutti *et al.* [241, 242]

IPolQ

Basically, the solvent reaction field about the ion is generated using MD simulation. The ion partial charges are then fitted by a RESP-like procedure to reproduce the ion dipole half way between those from the aqueous and vacuum phases. The fitted charges are then put back in another MD simulation in order to update the solvent reaction field. This procedure is repeated until the charges converge.

The LJ parameters are then adjusted to reproduce the hydration free energy of the ions. We constrain ourselves to vary only σ while keeping ε fixed. After this step, the ion with the updated LJ parameters will be subjected to the last solvent consistent field update above to get the final partial charges.

Briefly, each ion was soaked to an octahedral box of water and counter ions (either Na^+ or Cl^-) to make the system zero net charge. The box was large enough to separate the ion at least 20 Å from the edge. After equilibrating the system, an initial 5-ns MD was run at 450K in NPT ensemble, with PME, a 10 Å cutoff and 2 fs time step. Snapshots of the initial MD were collected every 250ps and were first minimized with 10.0 kcal/mol.$Å^2$ on all heavy atoms. Simulations of the fixed ion (while water molecules and counterions sampled around) were then carried out (for each snapshot) for 500ps in NVT ensemble at 298K. The locations of water molecules and counterions in these restraint-simulations were used to create a field of point charges surrounding the ion.

The mdgx code was then used to calculate the average electrostatic potential around the ion in both vacuum and condensed states. Essentially, water molecules and counterions (if any) within 5 Å of the ion were kept while the rest were discarded. To account for water molecules and counterions that are outside the 5 Å region, mdgx used 3 additional shells of point charges which were 5, 5.5 and 6 Å from the ion. Point charges in these shells were fitted in the same manner with RESP in order to approximate the influence of water molecules and counterions beyond the 5 Å cutoff. QM calculations were then performed at MP2/cc-pvTZ level in both vacuum and solvated states to obtain the electrostatic potential around the ion.

With these two electrostatic potentials in vacuum and solvated states, it was then possible to fit the atomic partial charges of the ion to reproduce the average of these two potentials. The charge fitting was carried out with some constraints that kept charges on equivalent atoms being equal and applied a harmonic restraint on methyl groups to keep their atomic charges small.

HFE calculation

Hydration free energies were computed using thermodynamic integration (TI). The ion was placed in a box of water molecules (with at least 20 Å to the box edge). After an initial minimization and 500 ps of NPT equilibration, TI was performed to compute HFE at 298.15K. The HFE calculation was divided into two steps: removing the partial atomic charges (*remQ*) and then removing the LJ interactions (*remLJ*). The *remQ* step was integrated via 5-abscissa Gaussian quadrature, while 12-abscissa was used for the *remLJ* step. Dynamics in each window were propagated for 4 ns at 298.15K, with PME, 10 Å cutoff and a 2 fs time step.

The ion HFEs were calculated as following:

$$\begin{aligned} \triangle G_{sim} &= \triangle G_{remQ_{(gas)}} - \triangle G_{remQ_{(sol)}} - \triangle G_{remLJ_{(sol)}} \\ &= \Delta G_{polar} + \Delta G_{apolar} \end{aligned} \tag{6.6}$$

where $\Delta G_{polar} = \triangle G_{remQ_{(gas)}} - \triangle G_{remQ_{(sol)}}$ and $\Delta G_{apolar} = -\triangle G_{remLJ_{(sol)}}$.

Free energy correction

Those corrections are needed so that the HFE computed by TI is comparable with those from the experiment.

- Correction for ion-ion self interaction as done in Refs. [259, 260]. The self-interaction potential (Wigner potential) of the ion is:

$$\phi_W\left(L\right) = \frac{q}{4\pi\epsilon_o\varepsilon}\frac{\xi}{L} \tag{6.7}$$

  where ξ is the electrostatic potential in a Wigner lattice at a charge site owing to the lattice images and the neutralizing background and varies inversely with the length of the cube L. The free energy correction is therefore:

$$\Delta G_{\phi_W} = \frac{q^2}{8\pi\varepsilon_o}\left(1 - \frac{1}{\varepsilon}\right)\frac{\xi}{L} \tag{6.8}$$

  This free energy correction has been already applied in Amber PME implementation.

- Correction for periodic ion-solvent interaction since solvent molecules far away from the ion respond to periodic images of the ion. This correction is always positive since we liberate the water molecules from the artificial interaction with the ion images. [261]

$$\Delta G = \frac{q^2}{6\epsilon_o L}\left(1 - \frac{1}{\epsilon}\right)\left[\left(\frac{R}{L}\right)^2 - \frac{4\pi}{15}\left(\frac{R}{L}\right)^5\right] \tag{6.9}$$

  where R is the radius of the ion. Several studies show this correction is usually small and thus will be neglected here.

- Correction for long range van der Waals interaction (as done in Shirts *et al.* [262]). This is also included automatically in Amber.

- Experimental data are usually obtained under standard atmospheric pressure, thus one also needs to convert the computed HFE to standard pressure by applying:

$$\begin{aligned}\Delta G^{*\rightarrow o} &= -RT\ln\left(\frac{V^o}{V^*}\right) \\ &= -1.893\, kcal\, mol^{-1}\end{aligned} \tag{6.10}$$

where the star and circle denote standard states of 1M and 1atm, respectively (as in Ref. [263]).

- By definition, in the HFE measurement the ion has to cross the air/water interface and the potential drop at this interface contributes to the free energy. In simulation, no explicit liquid/air boundary exists in the system, and therefore that effect is not included. The associated free energy contribution of the interface is:

  $$\Delta G_{surf} = q\phi \tag{6.11}$$

  where ϕ is the interfacial potential jump when transferring a point charge q from vacuum to aqueous phase. Following Warren and Patel [264] and Horinek *et al.* [265], the surface potential is chosen $\phi = -528mV$ and the HFE correction is $\Delta G_{surf} = -12.15q$ (kcal mol^{-1}). (See a more in-depth discussion about the effect of air-water interface to HFE in Ref. [266].)

The computed HFE in Amber TI is then corrected by:

$$\Delta G_{hyd} = \Delta G_{sim} + \Delta G^{*\to o} + \Delta G_{surf} \tag{6.12}$$

Experimental data on single ion hydration

Since experimental data of ion HFE requires extra thermodynamic assumptions, there are several set of data which are usually not compatible with each other. For example the difference between Na$^+$ HFE from Marcus and Tissandier *et al.* is about 16 kcal/mol while for Cl$^-$ the difference is about -7 kcal/mol, in the opposite direction (!!!). [3, 6] Warren and Patel showed that the reason behind these conflicts was the difference of the proton reference free energy. [264] Once all set are offset to the same absolute HFE of proton, all the experimental data become consistent (see Table 2 in Warren and Patel).

In this work, we choose the proton HFE to be -265.9 kcal/mol as in Tissandier *et al.* which was estimated by using the cluster-pair approximation and was supported by a lot of similar works thereafter. [2, 267] All the ion HFE data from Marcus are offset to the new reference proton HFE (see Table 6.1, the orig. and shifted Marcus columns). Also as pointed out by Warren and Patel, Marcus erroneously applied a correction of +1.893 kcal

Ion	Orig. Marcus	Shifted Marcus	Real ΔG^0	Intrinsic ΔG^0
F^-	-111.1	-99.9	-103.7	-115.9
Cl^-	-81.3	-70.1	-73.9	-86.0
Br^-	-75.3	-64.1	-67.9	-80.0
I^-	-65.7	-54.5	-58.3	-70.5
Li^+	-113.5	-124.7	-128.5	-116.4
Na^+	-87.2	-98.4	-102.2	-90.1
K^+	-70.5	-81.7	-85.5	-73.4
Rb^+	-65.7	-76.9	-80.7	-68.6
Cs^+	-59.8	-71.0	-74.7	-62.6
OH^-	-102.8	-91.6	-95.4	-107.5
SH^-	-70.5	-59.3	-63.1	-75.2
AcO^-	-87.2	-76.0	-79.8	-92.0
$HCOO^-$	-94.4	-83.2	-87.0	-99.1
NO_3^-	-71.7	-60.5	-64.3	-76.4
ClO_3^-	-66.9	-55.7	-59.5	-71.7
ClO_4^-	-49.0	-37.8	-41.6	-53.7
HCO_3^-	-80.1	-68.9	-72.7	-84.8
HSO_3^-	-55.9	-44.7	-48.5	-60.7
SCN^-	-66.9	-55.7	-59.5	-71.7
$H_2PO_4^-$	-111.1	-99.9	-103.7	-115.9
CO_3^{2-}	-314.3	-303.1	-306.9	-331.2
SO_4^{2-}	-258.1	-246.9	-250.7	-275.0
PO_4^{3-}	-660.9	-649.7	-653.4	-689.9
NH_4^+	-68.1	-79.3	-83.1	-71.0
$(CH_3)_4N^+$	-38.2	-49.4	-53.2	-41.1

Table 6.1: Corrected Marcus's hydration free energies of ions (kcal/mol). See the text above for the description of each column. Data in the "Orig. Marcus column" taken from Ref. [3].

mol^{-1} when converting from a standard state of 1 atm to 1M rather than -1.893 kcal mol^{-1} as in Eq. 6.10. Thus we apply a double correction to bring those numbers back to their actual values (column Real ΔG^0 in Table 6.1). Doing so we note that the HFE of monovalent ions are in greater improvement with other ion set. The intrinsic ΔG^0 is the HFE without the contribution from the interfacial potential jump.

For some more familiar ions, there are more updated data, for example from Truhlar and colleagues. [2] Thus it is reasonable to choose from those instead of from Marcus. We decide to use the data from Kelly *et al.* (which uses the cluster pair approximation for a very large set) for the following ions: $HCOO^-$, AcO^-, OH^-, HS^- and NH_4^+ (see Table 6.2). All the intrinsic HFEs will be the target for the calculated free energies to match.

Ion	HCOO$^-$	AcO$^-$	OH$^-$	HS$^-$	NH_4^+
Intrinsic ΔG^0	-88.4	-89.8	-116.9	-84.3	-73.1

Table 6.2: Intrinsic ΔG^0 (kcal/mol) of some more popular ions from Kelly *et al.* [2] with the interfacial correction -12.15q kcal/mol. The value of OH$^-$ from Tissandier *et al.* [6] is -115.1 kcal/mol.

References

[1] Murphy, L. R.; Matubayasi, N.; Payne, V. A.; Levy, R. M. *Folding and Design* **1998,** *3,* 105–118.

[2] Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2006,** *110,* 16066–16081.

[3] Marcus, Y. *Ion properties;* Marcel Dekker: New York, 1997.

[4] Reif, M. M.; Hunenberger, P. H.; Oostenbrink, C. *J. Chem. Theory Comput.* **2012,** *8,* 3705–3723.

[5] Steinbrecher, T.; Latzer, J.; Case, D. A. *J. Chem. Theory Comput.* **2012,** *8,* 4405–4412.

[6] Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R. *J. Phys. Chem. A* **1998,** *102,* 7787–7794.

[7] Bai, Y.; Greenfeld, M.; Travers, K. J.; Chu, V. B.; Lipfert, J.; Doniach, S.; Herschlag, D. *J. Am. Chem. Soc.* **2007,** *129,* 14981–14988.

[8] Imai, T.; Hiraoka, R.; Kovalenko, A.; Hirata, F. *J. Am. Chem. Soc.* **2005,** *127,* 15334–15335.

[9] Giambasu, G. M.; Luchko, T.; Herschlag, D.; York, D. M.; Case, D. A. *Biophys. J.* **2014,** *106,* 883–894.

[10] Blanchet, C. E.; Svergun, D. I. *Annu. Rev. Phys. Chem.* **2013,** *64,* 37–54.

[11] Park, S.; Bardhan, J. P.; Roux, B.; Makowski, L. *J. Chem. Phys.* **2009,** *130,* 134114.

[12] Nguyen, H. T.; Pabit, S. A.; Meisburger, S. P.; Pollack, L.; Case, D. A. *J. Chem. Phys.* **2014,** *141,* 22D508.

[13] Meisburger, S. P.; Pabit, S. A.; Pollack, L. *Biophys. J.* **2015,** *108,* 2886–2895.

[14] Ploetz, E. A.; Smith, P. E. *J. Phys. Chem. B* **2014,** *118,* 12844–12854.

[15] Richards, F. M. *Annu. Rev. Biophys. Bioeng.* **1977,** *6,* 151–176.

[16] Voss, N. R.; Gerstein, M. *Nucl. Acids Res.* **2010,** *38,* W555–W562.

[17] Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *J. Solution Chem.* **1981,** *10,* 563–595.

[18] Gianni, P.; Lepori, L. *J. Solution Chem.* **1996,** *25,* 1–42.

[19] Lepori, L.; Gianni, P. *J. Solution Chem.* **2000,** *29,* 405–447.

[20] Klibanov, A. M. *Trends Biochem. Sci.* **1989,** *14,* 141–144.

[21] Kurkal, V.; Daniel, R. M.; Finney, J. L.; Tehei, M.; Dunn, R. V.; Smith, J. C. *Biophys. J.* **2005,** *89,* 1282–1287.

[22] Israelachvili, J.; Wennerstrom, H. *Nature* **1996,** *379,* 219–225.

[23] Bagchi, B. *Chem. Rev.* **2005,** *105,* 3197–3219.

[24] Dill, K. A.; Truskett, T. M.; Vlachy, V.; Hribar-Lee, B. *Annu. Rev. Biophys. Biomol. Struct.* **2005,** *34,* 173–199.

[25] Wiggins, P. M. *Microbiol. Rev.* **1990,** *54,* 432–449.

[26] Levy, Y.; Onuchic, J. N. *Annu. Rev. Biophys. Biomol. Struct.* **2006,** *35,* 389–415.

[27] Makarov, V.; Pettitt, B. M.; Feig, M. *Acc. Chem. Res.* **2002,** *35,* 376–384.

[28] Ball, P. *Chem. Rev.* **2008,** *108,* 74–108.

[29] Ladbury, J. E. *Chemistry & Biology* **1996,** *3,* 973–980.

[30] Pal, S. K.; Zewail, A. H. *Chem. Rev.* **2004,** *104,* 2099–2124.

[31] Kuntz, I. D.; Kauzmann, W. *Adv. Protein Chem.* **1974,** *28,* 239–345.

[32] Halle, B. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **2004,** *359,* 1207–1224.

[33] Fogarty, A. C.; Duboue-Dijon, E.; Sterpone, F.; Hynes, J. T.; Laage, D. *Chem. Soc. Rev.* **2013,** *42,* 5672.

[34] Yu, B.; Blaber, M.; Gronenborn, A. M.; Clore, G. M.; Caspar, D. L. D. *Proc. Natl. Acad. Sci.* **1999,** *96,* 103–108.

[35] Williams, M. A.; Goodfellow, J. M.; Thornton, J. M. *Protein Sci.* **1994,** *3,* 1224–1235.

[36] Wlodawer, A.; Nachman, J.; Gilliland, G. L.; Gallagher, W.; Woodward, C. *J. Mol. Biol.* **1987,** *198,* 469–480.

[37] Burling, F. T.; Weis, W. I.; Flaherty, K. M.; Brunger, A. T. *Science* **1996,** *271,* 72–77.

[38] Svergun, D. I.; Richard, S.; Koch, M. H. J.; Sayers, Z.; Kuprin, S.; Zaccai, G. *Proc. Natl. Acad. Sci.* **1998,** *95,* 2267–2272.

[39] Merzel, F.; Smith, J. C. *Proc. Natl. Acad. Sci.* **2002,** *99,* 5378–5383.

[40] Makarov, V. A.; Andrews, B. K.; Smith, P. E.; Pettitt, B. M. *Biophys. J.* **2000,** *79,* 2966–2974.

[41] Hansen, J.-P.; McDonald, I. R. *Theory of simple liquids;* Elsevier Academic Press: 2006.

[42] Hummer, G.; Garcia, A. E.; Soumpasis, D. M. *Biophysical Journal* **1995,** *68,* 1639–1652.

[43] Hummer, G.; Garcia, A. E.; Soumpasis, D. M. *Faraday Discuss.* **1996,** 175–189.

[44] Koehl, P.; Delarue, M. *J. Chem. Phys.* **2010,** *132,* 064101.

[45] Fennell, C. J.; Kehoe, C. W.; Dill, K. A. *Proc. Natl. Acad. Sci.* **2011,** *108,* 3234–3239.

[46] Evans, R. *Advances in Physics* **1979,** *28,* 143–200.

[47] Lowen, H. *J. Phys.: Condens. Matter* **2002,** *14,* 11897.

[48] Wu, J.; Li, Z. *Annu. Rev. Phys. Chem.* **2007,** *58,* 85–112.

[49] Jeanmairet, G.; Levesque, M.; Vuilleumier, R.; Borgis, D. *J. Phys. Chem. Lett.* **2013,** *4,* 619–624.

[50] Guillot, B. *J. Mol. Liquids* **2002,** *101,* 219–260.

[51] Vega, C.; Abascal, J. L. F. *Phys. Chem. Chem. Phys.* **2011,** *13,* 19663.

[52] Wang, L.-P.; Head-Gordon, T.; Ponder, J. W.; Ren, P.; Chodera, J. D.; Eastman, P. K.; Martinez, T. J.; Pande, V. S. *J. Phys. Chem. B* **2013,** *117,* 9956–9972.

[53] Izadi, S.; Anandakrishnan, R.; Onufriev, A. V. *J. Phys. Chem. Lett.* **2014,** *5,* 3863–3871.

[54] Sokhan, V. P.; Jones, A. P.; Cipcigan, F. S.; Crain, J.; Martyna, G. J. *Proc. Natl. Acad. Sci.* **2015,** *112,* 6341–6346.

[55] Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. *J. Phys. Chem. B* **2015,** *119,* 5113–5123.

[56] Best, R. B.; Mittal, J. *J. Phys. Chem. B* **2010,** *114,* 14916–14923.

[57] Paschek, D.; Day, R.; Garcia, A. E. *Phys. Chem. Chem. Phys.* **2011,** *13,* 19840.

[58] Nerenberg, P. S.; Head-Gordon, T. *J. Chem. Theory Comput.* **2011,** *7,* 1220–1230.

[59] Best, R. B.; Zheng, W.; Mittal, J. *J. Chem. Theory Comput.* **2014,** *10,* 5113–5124.

[60] Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; de Groot, B. L.; Grubmuller, H. *J. Chem. Theory Comput.* **2015,** *11,* 5513–5524.

[61] Mercadante, D.; Milles, S.; Fuertes, G.; Svergun, D. I.; Lemke, E. A.; Grater, F. *J. Phys. Chem. B* **2015,** *119,* 7975–7984.

[62] Henriques, J.; Cragnell, C.; Skepo, M. *J. Chem. Theory Comput.* **2015,** *11,* 3420–3431.

[63] Chen, A. A.; Garcia, A. E. *Proc. Natl. Acad. Sci.* **2013,** *110,* 16820–16825.

[64] Bergonzo, C.; III, T. E. C. *J. Chem. Theory Comput.* **2015,** *11,* 3969–3972.

[65] Sponer, J.; Banas, P.; Jurecka, P.; Zgarbova, M.; Kuhrova, P.; Havrila, M.; Krepl, M.; Stadlbauer, P.; Otyepka, M. *J. Phys. Chem. Lett.* **2014,** *5,* 1771–1782.

[66] Page, M. J.; Cera, E. D. *Physiol. Rev.* **2006,** *86,* 1049–1092.

[67] Anderson, C. F.; Record, M. T. *Annu. Rev. Phys. Chem.* **1995,** *46,* 657–700.

[68] Draper, D. E. *RNA* **2004,** *10,* 335–343.

[69] Record, M. T.; Anderson, C. F.; Lohman, T. M. *Quart. Rev. Biophys.* **1978,** *11,* 103–178.

[70] Lipfert, J.; Doniach, S.; Das, R.; Herschlag, D. *Annu. Rev. Biochem.* **2014,** *83,* 813–841.

[71] Bloomfield, V. A. *Biopolymers* **1997,** *44,* 269–282.

[72] Wong, G. C. L.; Pollack, L. *Annu. Rev. Phys. Chem.* **2010,** *61,* 171–189.

[73] Grosberg, A. Y.; Nguyen, T. T.; Shklovskii, B. I. *Rev. Mod. Phys.* **2002,** *74,* 329–345.

[74] Sharp, K. A.; Honig, B. *Curr. Opin. Struct. Biol.* **1995,** *5,* 323–328.

[75] Jayaram, B.; Beveridge, D. L. *Annu. Rev. Biophys. Biomol. Struct.* **1996,** *25,* 367–394.

[76] Draper, D. E.; Grilley, D.; Soto, A. M. *Annu. Rev. Biophys. Biomol. Struct.* **2005,** *34,* 221–243.

[77] Kirkwood, J. G. *J. Chem. Phys.* **1934,** *2,* 767–781.

[78] Grochowski, P.; Trylska, J. *Biopolymers* **2008,** *89,* 93–113.

[79] Honig, B.; Nicholls, A. *Science* **1995,** *268,* 1144–1149.

[80] Warshel, A.; Russell, S. T. *Quart. Rev. Biophys.* **1984,** *17,* 283–422.

[81] Davis, M. E.; McCammon, J. A. *Chem. Rev.* **1990,** *90,* 509–521.

[82] Fogolari, F.; Brigo, A.; Molinari, H. *J. Mol. Recognit.* **2002,** *15,* 377–392.

[83] Sharp, K. A.; Honig, B. *Annu. Rev. Biophys. Biophys. Chem.* **1990,** *19,* 301–332.

[84] Manning, G. S. *J. Chem. Phys.* **1969,** *51,* 924–933.

[85] Manning, G. S. *Quart. Rev. Biophys.* **1978,** *11,* 179–246.

[86] Warwicker, J.; Watson, H. C. *J. Mol. Biol.* **1982,** *157,* 671–679.

[87] Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Natl. Acad. Sci.* **2001,** *98,* 10037–10041.

[88] Jayaram, B.; Sharp, K. A.; Honig, B. *Biopolymers* **1989,** *28,* 975–993.

[89] Pullman, A.; Pullman, B. *Quart. Rev. Biophys.* **1981,** *14,* 289–380.

[90] Denisov, V. P.; Halle, B. *Proc. Natl. Acad. Sci.* **2000,** *97,* 629–633.

[91] Tereshko, V.; Minasov, G.; Egli, M. *J. Am. Chem. Soc.* **1999,** *121,* 3590–3595.

[92] Howerton, S. B.; Sines, C. C.; VanDerveer, D.; Williams, L. D. *Biochemistry* **2001,** *40,* 10023–10031.

[93] Hud, N. V.; Sklenar, V.; Feigon, J. *J. Mol. Biol.* **1999,** *286,* 651–660.

[94] Ponomarev, S. Y.; Thayer, K. M.; Beveridge, D. L. *Proc. Natl. Acad. Sci.* **2004,** *101,* 14771–14775.

[95] Varnai, P.; Zakrzewska, K. *Nucl. Acids Res.* **2004,** *32,* 4269–4280.

[96] Rueda, M.; Cubero, E.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2004,** *87,* 800–811.

[97] Savelyev, A.; Papoian, G. A. *J. Am. Chem. Soc.* **2006,** *128,* 14506–14518.

[98] Kirmizialtin, S.; Elber, R. *J. Phys. Chem. B* **2010,** *114,* 8207–8220.

[99] Yoo, J.; Aksimentiev, A. *J. Phys. Chem. B* **2012,** *116,* 12946–12954.

[100] Savelyev, A.; MacKerell, A. D. *J. Phys. Chem. B* **2015,** *119,* 4428–4440.

[101] Pasi, M.; Maddocks, J. H.; Lavery, R. *Nucl. Acids Res.* **2015,** *43,* 2412–2423.

[102] Cate, J. H.; Gooding, A. R.; Podell, E.; Zhou, K.; Golden, B. L.; Kundrot, C. E.; Cech, T. R.; Doudna, J. A. *Science* **1996,** *273,* 1678–1685.

[103] Subirana, J. A.; Soler-Lopez, M. *Annu. Rev. Biophys. Biomol. Struct.* **2003,** *32,* 27–45.

[104] Bleam, M. L.; Anderson, C. F.; Record, M. T. *Proc. Natl. Acad. Sci.* **1980,** *77,* 3085–3089.

[105] Grilley, D.; Soto, A. M.; Draper, D. E. Chapter 3 - Direct quantitation of Mg2+–RNA Interactions by Use of a Fluorescent Dye. In *Meth. Enzymol.*, Vol. 455; Enzymology, B. M. i., Ed.; Academic Press: 2009.

[106] Pabit, S. A.; Meisburger, S. P.; Li, L.; Blose, J. M.; Jones, C. D.; Pollack, L. *J. Am. Chem. Soc.* **2010,** *132,* 16334–16336.

[107] Barker, J. A.; Henderson, D. *Rev. Mod. Phys.* **1976,** *48,* 587–671.

[108] Chandler, D.; Andersen, H. C. *J. Chem. Phys.* **1972,** *57,* 1930–1937.

[109] Hirata, F.; Rossky, P. J. *Chem. Phys. Lett.* **1981,** *83,* 329–334.

[110] Perkyns, J. S.; Montgomery Pettitt, B. *Chem. Phys. Lett.* **1992,** *190,* 626–630.

[111] Beglov, D.; Roux, B. *J. Chem. Phys.* **1995,** *103,* 360–364.

[112] Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997,** *101,* 7821–7826.

[113] Ikeguchi, M.; Doi, J. *J. Chem. Phys.* **1995,** *103,* 5011–5017.

[114] Cortis, C. M.; Rossky, P. J.; Friesner, R. A. *J. Chem. Phys.* **1997,** *107,* 6400–6414.

[115] Kovalenko, A.; Hirata, F. *Chem. Phys. Lett.* **1998,** *290,* 237–244.

[116] Luchko, T.; Joung, I. S.; Case, D. A. Chapter 4 Integral Equation Theory of Biomolecules and Electrolytes. In *Innovations in Biomolecular Modeling and Simulations*; 2012.

[117] Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V. *Chem. Rev.* **2015,** *115,* 6312–6356.

[118] Miyata, T.; Hirata, F. *J. Comput. Chem.* **2008,** *29,* 871–882.

[119] Luchko, T.; Gusarov, S.; Roe, D. R.; Simmerling, C.; Case, D. A.; Tuszynski, J.; Kovalenko, A. *J. Chem. Theory Comput.* **2010,** *6,* 607–624.

[120] Perkyns, J.; Pettitt, B. M. *J. Chem. Phys.* **1992,** *97,* 7656–7666.

[121] Joung, I. S.; Luchko, T.; Case, D. A. *J. Chem. Phys.* **2013,** *138,*.

[122] Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **1999,** *110,* 10095–10112.

[123] Kast, S. M.; Kloss, T. *J. Chem. Phys.* **2008,** *129,* 236101.

[124] Hirata, F. Chapter 1 - Theory of molecular liquids. In *Molecular Theory of Solvation*; Kluwer academic publishers: 2004.

[125] Morita, T.; Hiroike, K. *Prog. Theor. Phys.* **1960,** *23,* 1003–1027.

[126] Palmer, D. S.; Frolov, A. I.; Ratkova, E. L.; Fedorov, M. V. *J. Phys.: Condens. Matter* **2010,** *22,* 492101.

[127] Truchon, J.-F.; Pettitt, B. M.; Labute, P. *J. Chem. Theory Comput.* **2014,** *10,* 934–941.

[128] Sergiievskyi, V.; Jeanmairet, G.; Levesque, M.; Borgis, D. *J. Chem. Phys.* **2015,** *143,* 184116.

[129] Yoshida, N.; Phongphanphanee, S.; Maruyama, Y.; Imai, T.; Hirata, F. *J. Am. Chem. Soc.* **2006,** *128,* 12042–12043.

[130] Phongphanphanee, S.; Rungrotmongkol, T.; Yoshida, N.; Hannongbua, S.; Hirata, F. *J. Am. Chem. Soc.* **2010,** *132,* 9782–9788.

[131] Phongphanphanee, S.; Yoshida, N.; Hirata, F. *J. Phys. Chem. B* **2010,** *114,* 7967–7973.

[132] Nikolic, D.; Blinov, N.; Wishart, D.; Kovalenko, A. *J. Chem. Theory Comput.* **2012,** *8,* 3356–3372.

[133] Stumpe, M. C.; Blinov, N.; Wishart, D.; Kovalenko, A.; Pande, V. S. *J. Phys. Chem. B* **2011,** *115,* 319–328.

[134] Howard, J. J.; Lynch, G. C.; Pettitt, B. M. *J. Phys. Chem. B* **2011,** *115,* 547–556.

[135] Svergun, D.; Barberato, C.; Koch, M. H. J. *J. Appl. Cryst.* **1995,** *28,* 768–773.

[136] Poitevin, F.; Orland, H.; Doniach, S.; Koehl, P.; Delarue, M. *Nucl. Acids Res.* **2011,** *39,* W184–W189.

[137] Virtanen, J. J.; Makowski, L.; Sosnick, T. R.; Freed, K. F. *Biophys. J.* **2011,** *101,* 2061–2069.

[138] Cromer, D. T.; Mann, J. B. *Acta Cryst. A* **1968,** *24,* 321–324.

[139] Su, Z.; Coppens, P. *Acta Cryst. A* **1997,** *53,* 749–762.

[140] Glatter, O.; Kratky, O. *Small Angle X-ray Scattering;* Academic Press: 1982.

[141] Feigin, L. A.; Svergun, D. I. *Structure Analysis by Small-Angle X-Ray and Neutron Scattering;* Springer US: Boston, MA, 1987.

[142] Koch, M. H. J.; Vachette, P.; Svergun, D. I. *Quart. Rev. Biophys.* **2003,** *36,* 147–227.

[143] Svergun, D. I.; Koch, M. H. J. *Rep. Prog. Phys.* **2003,** *66,* 1735.

[144] Putnam, C. D.; Hammel, M.; Hura, G. L.; Tainer, J. A. *Quart. Rev. Biophys.* **2007,** *40,* 191–285.

[145] Rambo, R. P.; Tainer, J. A. *Annu. Rev. Biophys.* **2013,** *42,* 415–441.

[146] Schneidman-Duhovny, D.; Hammel, M.; Sali, A. *Nucl. Acids Res.* **2010,** *38,* W540–W544.

[147] Schneidman-Duhovny, D.; Hammel, M.; Tainer, J. A.; Sali, A. *Biophys. J.* **2013,** *105,* 962–974.

[148] Grishaev, A.; Guo, L.; Irving, T.; Bax, A. *J. Am. Chem. Soc.* **2010,** *132,* 15484–15486.

[149] Azuara, C.; Orland, H.; Bon, M.; Koehl, P.; Delarue, M. *Biophys. J.* **2008,** *95,* 5587–5605.

[150] Makarov, V. A.; Andrews, B. K.; Pettitt, B. M. *Biopolymers* **1998,** *45,* 469–478.

[151] Virtanen, J. J.; Makowski, L.; Sosnick, T. R.; Freed, K. F. *Biophys. J.* **2010,** *99,* 1611–1619.

[152] Pavlov, M. Y.; Fedorov, B. A. *Biopolymers* **1983,** *22,* 1507–1522.

[153] Merzel, F.; Smith, J. C. *Acta Cryst. D* **2002,** *58,* 242–249.

[154] Seki, Y.; Tomizawa, T.; Khechinashvili, N. N.; Soda, K. *Biophys. Chem.* **2002,** *95,* 235–252.

[155] Oroguchi, T.; Hashimoto, H.; Shimizu, T.; Sato, M.; Ikeguchi, M. *Biophys. J.* **2009,** *96,* 2808–2822.

[156] Chen, P.-c.; Hub, J. S. *Biophys. J.* **2014,** *107,* 435–447.

[157] Berlin, K.; Gumerov, N. A.; Fushman, D.; Duraiswami, R. *J. Appl. Cryst.* **2014,** *47,* 755–761.

[158] Lebedev, V.; Laikov, D. *Doklady Mathematics* **1999,** *59,* 477–481.

[159] Treutler, O.; Ahlrichs, R. *J. Chem. Phys.* **1995,** *102,* 346–354.

[160] Zheng, G.; Lu, X.-J.; Olson, W. K. *Nucl. Acids Res.* **2009,** *37,* W240–W246.

[161] Fedoroff OYu, n.; Salazar, M.; Reid, B. R. *J. Mol. Biol.* **1993,** *233,* 509–523.

[162] Horton, N. C.; Finzel, B. C. *J. Mol. Biol.* **1996,** *264,* 521–533.

[163] Cheatham, T. E.; Kollman, P. A. *J. Am. Chem. Soc.* **1997,** *119,* 4805–4825.

[164] Romainczyk, O.; Endeward, B.; Prisner, T. F.; Engels, J. W. *Mol. BioSyst.* **2011,** *7,* 1050–1052.

[165] Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham III, T. E.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007,** *92,* 3817–3829.

[166] Zgarbova, M.; Luque, F. J.; Sponer, J.; Cheatham, T. E.; Otyepka, M.; Jurecka, P. *J. Chem. Theory Comput.* **2013,** *9,* 2339–2354.

[167] Krepl, M.; Zgarbova, M.; Stadlbauer, P.; Otyepka, M.; Banas, P.; Koca, J.; Cheatham, T. E.; Jurecka, P.; Sponer, J. *J. Chem. Theory Comput.* **2012,** *8,* 2506–2520.

[168] Zgarbova, M.; Otyepka, M.; Sponer, J.; Mladek, A.; Banas, P.; Cheatham, T. E.; Jurecka, P. *J. Chem. Theory Comput.* **2011,** *7,* 2886–2902.

[169] Sorin, E. J.; Rhee, Y. M.; Pande, V. S. *Biophys. J.* **2005,** *88,* 2516–2524.

[170] Besseova, I.; Banas, P.; Kuhrova, P.; Kosinova, P.; Otyepka, M.; Sponer, J. *J. Phys. Chem. B* **2012,** *116,* 9899–9916.

[171] Kuhrova, P.; Otyepka, M.; Sponer, J.; Banas, P. *J. Chem. Theory Comput.* **2014,** *10,* 401–411.

[172] Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987,** *91,* 6269–6271.

[173] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983,** *79,* 926–935.

[174] Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004,** *120,* 9665–9678.

[175] Joung, I. S.; Cheatham, T. E. *J. Phys. Chem. B* **2008,** *112,* 9020–9041.

[176] Hart, K.; Foloppe, N.; Baker, C. M.; Denning, E. J.; Nilsson, L.; MacKerell, A. D. *J. Chem. Theory Comput.* **2012,** *8,* 348–362.

[177] Denning, E. J.; Priyakumar, U. D.; Nilsson, L.; Mackerell, A. D. *J. Comput. Chem.* **2011,** *32,* 1929–1943.

[178] Gotz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. *J. Chem. Theory Comput.* **2012,** *8,* 1542–1555.

[179] Salomon-Ferrer, R.; Gotz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. *J. Chem. Theory Comput.* **2013,** *9,* 3878–3888.

[180] Le Grand, S.; Gotz, A. W.; Walker, R. C. *Comp. Phys. Comm.* **2013,** *184,* 374–380.

[181] Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993,** *98,* 10089–10092.

[182] Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995,** *103,* 8577–8593.

[183] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977,** *23,* 327–341.

[184] Nguyen, H.; Perez, A.; Bermeo, S.; Simmerling, C. *J. Chem. Theory Comput.* **2015,** *11,* 3714–3728.

[185] Shahrokh, K.; Orendt, A.; Yost, G. S.; Cheatham, T. E. *J. Comput. Chem.* **2012,** *33,* 119–133.

[186] Li, P.; Roberts, B. P.; Chakravorty, D. K.; Merz, K. M. *J. Chem. Theory Comput.* **2013,** *9,* 2733–2748.

[187] Kovalenko, A.; Ten-no, S.; Hirata, F. *J. Comput. Chem.* **1999,** *20,* 928–936.

[188] Tiede, D. M.; Zhang, R.; Seifert, S. *Biochemistry* **2002,** *41,* 6605–6614.

[189] Bardhan, J.; Park, S.; Makowski, L. *J. Appl. Cryst.* **2009,** *42,* 932–943.

[190] Moore, P. B. *Biophys. J.* **2014,** *106,* 1489–1496.

[191] Das, R.; Mills, T. T.; Kwok, L. W.; Maskel, G. S.; Millett, I. S.; Doniach, S.; Finkelstein, K. D.; Herschlag, D.; Pollack, L. *Phys. Rev. Lett.* **2003,** *90,* 188103.

[192] Chu, V. B.; Bai, Y.; Lipfert, J.; Herschlag, D.; Doniach, S. *Biophys. J.* **2007,** *93,* 3202–3209.

[193] Kirmizialtin, S.; Silalahi, A. R. J.; Elber, R.; Fenley, M. O. *Biophys. J.* **2012,** *102,* 829–838.

[194] Feig, M.; Pettitt, B. M. *Biophys. J.* **1999,** *77,* 1769–1781.

[195] Kirmizialtin, S.; Pabit, S. A.; Meisburger, S. P.; Pollack, L.; Elber, R. *Biophys. J.* **2012,** *102,* 819–828.

[196] Kirkwood, J. G.; Buff, F. P. *J. Chem. Phys.* **1951,** *19,* 774–777.

[197] Svergun, D. I. *J. Appl. Cryst.* **1992,** *25,* 495–503.

[198] Chen, P.-c.; Hub, J. S. *Biophys. J.* **2015,** *108,* 2573–2584.

[199] Kimanius, D.; Pettersson, I.; Schluckebier, G.; Lindahl, E.; Andersson, M. *J. Chem. Theory Comput.* **2015,** *11,* 3491–3498.

[200] Zhong, D.; Pal, S. K.; Zewail, A. H. *Chem. Phys. Lett.* **2011,** *503,* 1–11.

[201] Shui, X.; McFail-Isom, L.; Hu, G. G.; Williams, L. D. *Biochemistry* **1998,** *37,* 8341–8355.

[202] Nakano, S.-i.; Fujimoto, M.; Hara, H.; Sugimoto, N. *Nucl. Acids Res.* **1999,** *27,* 2957–2965.

[203] Strauss, U. P.; Helfgott, C.; Pink, H. *J. Phys. Chem.* **1967,** *71,* 2550–2556.

[204] Dingenouts, N.; Patel, M.; Rosenfeldt, S.; Pontoni, D.; Narayanan, T.; Ballauff, M. *Macromolecules* **2004,** *37,* 8152–8159.

[205] Pollack, L. *Annu. Rev. Biophys.* **2011,** *40,* 225–242.

[206] Guilleaume, B.; Ballauff, M.; Goerigk, G.; Wittemann, M.; Rehahn, M. *Colloid Polym. Sci.* **2001,** *279,* 829–835.

[207] Dingenouts, N.; Merkle, R.; Guo, X.; Narayanan, T.; Goerigk, G.; Ballauff, M. *J. Appl. Cryst.* **2003,** *36,* 578–582.

[208] Moore, P. B. *J. Appl. Cryst.* **1980,** *13,* 168–175.

[209] Patel, M.; Rosenfeldt, S.; Ballauff, M.; Dingenouts, N.; Pontoni, D.; Narayanan, T. *Phys. Chem. Chem. Phys.* **2004,** *6,* 2962–2967.

[210] Pabit, S. A.; Finkelstein, K. D.; Pollack, L. Chapter 19 - Using Anomalous Small Angle X-Ray Scattering to Probe the Ion Atmosphere Around Nucleic Acids. In *Methods in Enzymology*, Vol. Volume 469; Herschlag, D., Ed.; Academic Press: 2009.

[211] Gebala, M.; Giambasu, G. M.; Lipfert, J.; Bisaria, N.; Bonilla, S.; Li, G.; York, D. M.; Herschlag, D. *J. Am. Chem. Soc.* **2015,** *137,* 14705–14715.

[212] Glatter, O. *J. Appl. Cryst.* **1977,** *10,* 415–421.

[213] Svergun, D. I.; Semenyuk, A. V.; Feigin, L. A. *Acta. Cryst. A* **1988,** *44,* 244–250.

[214] Hansen, S. *J. Appl. Cryst.* **2000,** *33,* 1415–1421.

[215] Hansen, S. Bayesian Methods in SAXS and SANS Structure Determination. In *Bayesian Methods in Structural Bioinformatics*; Hamelryck, T.; Mardia, K.; Ferkinghoff-Borg, J., Eds.; Statistics for Biology and Health Springer Berlin Heidelberg: 2012.

[216] Hansen, S. *J. Appl. Cryst.* **2012,** *45,* 566–567.

[217] Glatter, O. *J. Appl. Cryst.* **1979,** *12,* 166–175.

[218] Glatter, O.; Hainisch, B. *J. Appl. Cryst.* **1984,** *17,* 435–441.

[219] Schnieders, M. J.; Fenn, T. D.; Pande, V. S.; Brunger, A. T. *Acta Cryst. D* **2009,** *65,* 952–965.

[220] Agarwal, R. C. *Acta Cryst. A* **1978,** *34,* 791–809.

[221] Afonine, P. V.; Urzhumtsev, A. *Acta Cryst. A* **2004,** *60,* 19–32.

[222] Stillinger, F. H. *J. Solution Chem.* **1973,** *2,* 141–158.

[223] Edward, J. T.; Farrell, P. G. *Can. J. Chem.* **1975,** *53,* 2965–2970.

[224] Patel, N.; Dubins, D. N.; Pomes, R.; Chalikian, T. V. *Biophys. Chem.* **2012,** *161,* 46–49.

[225] Kofinger, J.; Hummer, G. *Phys. Rev. E* **2013,** *87,* 052712.

[226] Chiu, T. K.; Dickerson, R. E. *J. Mol. Biol.* **2000,** *301,* 915–945.

[227] Davey, C. A.; Richmond, T. J. *Proc. Natl. Acad. Sci.* **2002,** *99,* 11169–11174.

[228] Ahmad, R.; Arakawa, H.; Tajmir-Riahi, H. A. *Biophys. J.* **2003,** *84,* 2460–2466.

[229] Strick, R.; Strissel, P. L.; Gavrilov, K.; Levi-Setti, R. *J. Cell Biol.* **2001,** *155,* 899–910.

[230] Egli, M. *Chemistry & Biology* **2002,** *9,* 277–286.

[231] Giambasu, G. M.; Gebala, M. K.; Panteva, M. T.; Luchko, T.; Case, D. A.; York, D. M. *Nucl. Acids Res.* **2015,** *43,* 8405–8415.

[232] Li, P.; Merz, K. M. *J. Chem. Theory Comput.* **2014,** *10,* 289–297.

[233] Li, P.; Song, L. F.; Merz, K. M. *J. Phys. Chem. B* **2015,** *119,* 883–895.

[234] Pollack, L.; Tate, M. W.; Finnefrock, A. C.; Kalidas, C.; Trotter, S.; Darnton, N. C.; Lurio, L.; Austin, R. H.; Batt, C. A.; Gruner, S. M.; Mochrie, S. G. J. *Phys. Rev. Lett.* **2001,** *86,* 4962–4965.

[235] Cammarata, M.; Levantino, M.; Schotte, F.; Anfinrud, P. A.; Ewald, F.; Choi, J.; Cupane, A.; Wulff, M.; Ihee, H. *Nat. Meth.* **2008,** *5,* 881–886.

[236] Kim, J. G.; Kim, T. W.; Kim, J.; Ihee, H. *Acc. Chem. Res.* **2015,** *48,* 2200–2208.

[237] Gotz, A. W.; Bucher, D.; Lindert, S.; McCammon, J. A. *J. Chem. Theory Comput.* **2014,** *10,* 1631–1637.

[238] Savelyev, A.; MacKerell, A. D. *J. Phys. Chem. B* **2014,** *118,* 6742–6757.

[239] Ploetz, E. A.; Bentenitis, N.; Smith, P. E. *Fluid Phase Equilib.* **2010,** *290,* 43.

[240] Chapman, D. E.; Steck, J. K.; Nerenberg, P. S. *J. Chem. Theory Comput.* **2014,** *10,* 273–281.

[241] Cerutti, D. S.; Swope, W. C.; Rice, J. E.; Case, D. A. *J. Chem. Theory Comput.* **2014,** *10,* 4515–4534.

[242] Cerutti, D. S.; Rice, J. E.; Swope, W. C.; Case, D. A. *J. Phys. Chem. B* **2013,** *117,* 2328–2338.

[243] Kharakoz, D. P. *J. Solution Chem.* **1992,** *21,* 569–595.

[244] Harano, Y.; Imai, T.; Kovalenko, A.; Kinoshita, M.; Hirata, F. *J. Chem. Phys.* **2001,** *114,* 9506–9511.

[245] Imai, T.; Kovalenko, A.; Hirata, F. *J. Phys. Chem. B* **2005,** *109,* 6658–6665.

[246] DeVane, R.; Ridley, C.; Larsen, R. W.; Space, B.; Moore, P. B.; Chan, S. I. *Biophys. J.* **2003,** *85,* 2801–2807.

[247] Medvedev, N. N.; Voloshin, V. P.; Kim, A. V.; Anikeenko, A. V.; Geiger, A. *J. Struct. Chem.* **2014,** *54,* 271–288.

[248] Vilseck, J. Z.; Tirado-Rives, J.; Jorgensen, W. L. *Phys. Chem. Chem. Phys.* **2015,** *17,* 8407–8415.

[249] Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. *J. Chem. Theory Comput.* **2015,** *11,* 3696–3713.

[250] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. *J. Chem. Theory Comput.* **2012,** *8,* 3257–3273.

[251] Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. *J. Comput. Chem.* **2004,** *25,* 1656–1676.

[252] Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; Gunsteren, W. F. v. *Eur. Biophys. J.* **2011,** *40,* 843–856.

[253] Weerasinghe, S.; Gee, M. B.; Kang, M.; Bentenitis, N.; Smith, P. E. Developing Force Fields from the Microscopic Structure of Solutions: The Kirkwood–Buff Approach. In *Modeling Solvent Environments*; Feig, M., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: 2010.

[254] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996,** *118,* 11225–11236.

[255] Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. *J. Phys. Chem. B* **2010,** *114,* 2549–2564.

[256] Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. *J. Chem. Theory Comput.* **2013,** *9,* 4046–4063.

[257] Reif, M. M.; Winger, M.; Oostenbrink, C. *J. Chem. Theory Comput.* **2013,** *9,* 1247–1264.

[258] Bjorling, A.; Niebling, S.; Marcellini, M.; van der Spoel, D.; Westenhoff, S. *J. Chem. Theory Comput.* **2015,** *11,* 780–787.

[259] Hummer, G.; Pratt, L. R.; Garcia, A. E. *J. Phys. Chem.* **1996,** *100,* 1206–1215.

[260] Darden, T.; Pearlman, D.; Pedersen, L. G. *J. Chem. Phys.* **1998,** *109,* 10921–10935.

[261] Hummer, G.; Pratt, L. R.; Garcia, A. E. *J. Chem. Phys.* **1997,** *107,* 9275–9277.

[262] Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003,** *119,* 5740–5761.

[263] Marcus, Y. *J. Chem. Soc., Faraday Trans.* **1991,** *87,* 2995–2999.

[264] Warren, G. L.; Patel, S. *J. Chem. Phys.* **2007,** *127,* 064509.

[265] Horinek, D.; Mamatkulov, S. I.; Netz, R. R. *J. Chem. Phys.* **2009,** *130,* 124507.

[266] Beck, T. L. *Chem. Phys. Lett.* **2013,** *561-562,* 1–13.

[267] Camaioni, D. M.; Schwerdtfeger, C. A. *J. Phys. Chem. A* **2005,** *109,* 10795–10797.