Simulated x-ray scattering of protein solutions using explicit-solvent models

Sanghyun Park,[1,a)] Jaydeep P. Bardhan,[2] Benoît Roux,[2,3] and Lee Makowski[2]
[1]*Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA*
[2]*Biosciences Division, Argonne National Laboratory, Argonne, Illinois 60439, USA*
[3]*Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, Illinois 60637, USA*

(Received 29 December 2008; accepted 25 February 2009; published online 7 April 2009)

X-ray solution scattering shows new promise for the study of protein structures, complementing crystallography and nuclear magnetic resonance. In order to realize the full potential of solution scattering, it is necessary to not only improve experimental techniques but also develop accurate and efficient computational schemes to relate atomistic models to measurements. Previous computational methods, based on continuum models of water, have been unable to calculate scattering patterns accurately, especially in the wide-angle regime which contains most of the information on the secondary, tertiary, and quaternary structures. Here we present a novel formulation based on the atomistic description of water, in which scattering patterns are calculated from atomic coordinates of protein and water. Without any empirical adjustments, this method produces scattering patterns of unprecedented accuracy in the length scale between 5 and 100 Å, as we demonstrate by comparing simulated and observed scattering patterns for myoglobin and lysozyme. © *2009 American Institute of Physics*. [DOI: 10.1063/1.3099611]

I. INTRODUCTION

X-ray solution scattering, although a well-established technique, shows new promise for the study of structural properties of proteins.[1,2] Unlike in crystallography where protein motions are restricted by crystal contacts, proteins in solution are subject to natural dynamics. Solution scattering, therefore, has great potential for observing biologically important conformational changes and fluctuations in physiological conditions. Moreover, because only minimal preparations are required without the need for crystallization, solution scattering is an ideal technique for high-throughput research efforts. Accordingly, solution scattering is gaining popularity as a technique complementary to crystallography and nuclear magnetic resonance.

Based on the resolution of measurements, x-ray solution-scattering experiments are often categorized into small-angle x-ray scattering (SAXS) and wide-angle x-ray scattering (WAXS). Typically, SAXS probes protein structures between 50 and 1000 Å and provides information on size and shape.[1,2] WAXS explores structures around 2–100 Å and provides information on secondary, tertiary, and quaternary structures.[3,4]

A solution-scattering pattern represents an average over an ensemble of protein structures that are free to tumble and, therefore, does not contain sufficient information to reconstruct a full three-dimensional structure. Clearly, however, solution-scattering data contain significant amount of information that is useful for various studies of protein conformations.[2] In particular, it has been found that WAXS data can be used to characterize different protein folds.[5] The full information content of solution-scattering data is yet to be determined, and understanding the relationship between experiments and theoretical modeling is especially critical for extracting the maximum amount of information from solution-scattering data.

A fundamental component in theoretical modeling of solution scattering is the task of calculating scattering patterns from atomic coordinates. This is a highly nontrivial task because of the significant contribution of the solvent, water in most cases, in solution scattering. Virtually all the methods we are aware of, from the early work of Fraeser *et al.*[6] to the widely used software CRYSOL,[7] employ continuum models of water to account for the contribution of the solvent. Under such continuum models, scattering intensities are calculated in terms of the electron density of the protein and the volume of the solvent excluded by the protein.[6–8] Some methods include "solvation layers" in order to account for the higher density of water near the protein compared to the bulk.[7] Attempts have been made to include explicit solvent molecules in the solvation layer.[9] But even these methods still rely on the continuum description of water outside the solvation layer.

A continuum description of water should be reasonable at the 50 Å resolution, where the liquid has no apparent internal structure.[10] However, whether water can be treated as continuum at the 10 Å resolution and beyond is clearly more questionable. In that regard, it is not surprising that when the continuum-solvent approximations are extended to the resolution of WAXS, significant and seemingly systematic discrepancies are found between the observed and calculated scattering patterns.[11] Although consideration of chemical diversity of atoms on the protein surface can result in some improvement, a fundamental limit in the accuracy of

[a)]Electronic mail: sp.spark@gmail.com.

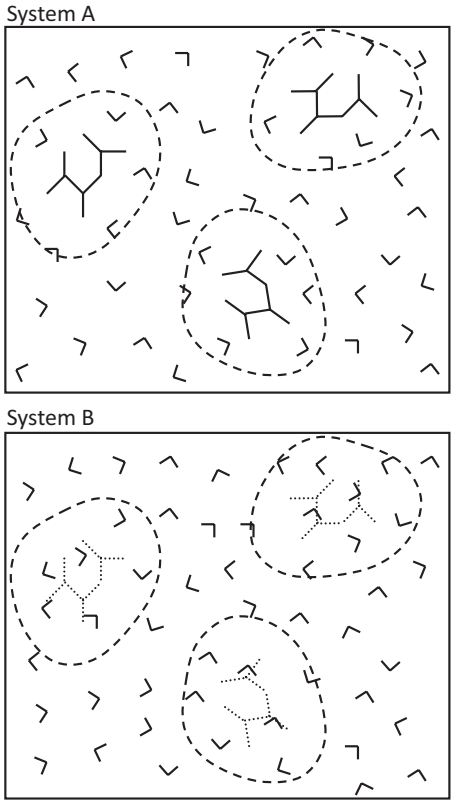

FIG. 1. Two systems for x-ray scattering. System A contains proteins in water, and system B contains only water, in the same volume. The dashed envelopes in system A represent water shells assigned to the protein molecules. The same envelopes are drawn in system B as if images (dotted lines) of the identical protein coordinates were present.

continuum-water methods remains that can be overcome only by adopting an atomistic description for water.

In this paper, we present a novel method for simulating x-ray scattering of protein solutions based entirely on the atomistic description of water, without resorting to the notion of excluded volume. Atomic coordinates of water around proteins are generally not available from experiments but can be generated from relatively short molecular dynamics (MD) simulations. Using myoglobin (Mb) and lysozyme (Lyz) as test cases, we show that the new method produces scattering patterns of high accuracy in both the SAXS and WAXS regimes. We identify the possible sources of the residual discrepancies between the observed and simulated data and discuss ways of further improvements.

II. THEORY

In a typical solution-scattering experiment, two sets of data are collected, one from a protein solution and another from a pure solvent of the same volume. Figure 1 schematically shows such two systems. Let $A(\mathbf{r})$ be the instantaneous electron density of system A (a protein solution), which implicitly depends on the atomic coordinates of the proteins and the solvent. The scattering intensity from system A is then given as

$$I_{\mathrm{A}}(q) = \langle |\tilde{A}(\mathbf{q})|^2 \rangle, \tag{1}$$

$$\tilde{A}(\mathbf{q}) = \int d\mathbf{r} e^{-i\mathbf{q}\cdot\mathbf{r}} A(\mathbf{r}), \tag{2}$$

where $\mathbf{q}$ is the scattering wave vector.[12] The magnitude of the scattering wave vector q is related to the scattering angle 2θ by $q=(4\pi/\lambda)\sin\theta$, where λ is the wavelength of the x-ray beam. The ensemble average $\langle\cdots\rangle$ denotes a thermal average over the protein and solvent degrees of freedom that include translations, rotations, and internal motions of the protein and solvent molecules. The rotational averaging makes I_{A} isotropic and dependent only on the magnitude of the scattering wave vector. Similarly, the scattering intensity from system B (pure solvent) is determined from its electron density $B(\mathbf{r})$:

$$I_{\mathrm{B}}(q) = \langle |\tilde{B}(\mathbf{q})^2| \rangle, \tag{3}$$

$$\tilde{B}(\mathbf{q}) = \int d\mathbf{r} e^{-i\mathbf{q}\cdot\mathbf{r}} B(\mathbf{r}). \tag{4}$$

A. Excess intensity

The main quantity of interest, which we wish to calculate from simulations, is the excess intensity:

$$\Delta I(q) = I_{\mathrm{A}}(q) - I_{\mathrm{B}}(q). \tag{5}$$

Using the excess intensity is a way of subtracting the background scattering. Different ways of background subtraction have been suggested, but the scheme corresponding to Eq. (5) seems most straightforward and least prone to uncertainties. A scaled background-subtraction scheme has been used before:

$$\Delta I_{\mathrm{scaled}}(q) = I_{\mathrm{A}}(q) - (1-f) I_{\mathrm{B}}(q), \tag{6}$$

where f is the volume fraction of the solution taken up by the proteins.[13] This scheme is motivated by the continuum-solvent picture, in which $\Delta I_{\mathrm{scaled}}$ can be interpreted as the scattering intensity from the excess electron density of the protein in the background of the *uniform* electron density of the solvent. This scheme requires estimation of f, which often introduces large uncertainties. The direct subtraction scheme of Eq. (5) seems more natural and painless for both the analyses of experimental data and the atomistic-water-based simulations. The choice between the two schemes is largely immaterial at small angles where I_{B} is negligible compared to I_{A}. At high angles, however, the magnitudes of the two intensities are similar, and it is important to use a consistent background-subtraction scheme when comparing observed and simulated scattering patterns.

Since experimental scattering intensities are measured in arbitrary units, observed intensities are proportional, rather than equal, to $I_{\mathrm{A}}(q)$ or $I_{\mathrm{B}}(q)$. The proportionality coefficient, however, can be kept constant by ensuring the same scattering volume (the volume of specimen that the x-ray beam goes through) in all measurements and by normalizing scattered intensities by the incoming beam intensity. The proportionality coefficient, then, can be estimated by comparing the observed and simulated excess intensities.

By decomposing the ensemble average into averages over the protein and solvent degrees of freedom, the excess intensity can be expressed as

$$\Delta I(q) = \langle D(\mathbf{q}) \rangle_p, \tag{7}$$

$$D(\mathbf{q}) \equiv \langle |\tilde{A}(\mathbf{q})|^2 \rangle_w - \langle |\tilde{B}(\mathbf{q})|^2 \rangle_w, \tag{8}$$

where $\langle \cdots \rangle_w$ and $\langle \cdots \rangle_p$ denote the ensemble averages over the solvent and protein degrees of freedom, respectively. $D(\mathbf{q})$ depends implicitly on the protein coordinates.

The protein degrees of freedom include translations, rotations, and internal motions of the protein molecules. As will be explained in Sec. II C, in this work we consider only rotational averages for proteins, which are calculated as integrals over the solid angle of $\mathbf{q}$ rather than in the form of ensemble averages. Hereafter, for the simplicity of notation, we use $\langle \cdots \rangle$ to denote an ensemble average over the *solvent* degrees of freedom.

B. Averaging over the solvent degrees of freedom

Let us first consider the calculation of $D(\mathbf{q})$ by averaging over the solvent degrees of freedom, with a given set of protein coordinates fixed in space. For system A, which contains N protein molecules, we assign a water shell to each protein (Fig. 1) and separate the instantaneous electron density $A(\mathbf{r})$ into

$$A(\mathbf{r}) = \sum_{n=1}^{N} A_n(\mathbf{r}) + A_0(\mathbf{r}), \tag{9}$$

where $A_n(\mathbf{r})$ for $n=1,\ldots,N$ is the electron density of protein n plus its water shell and $A_0(\mathbf{r})$ is the electron density of the solvent that is not included in any of the water shells. We define the water shells by the following proximity criterion: a water molecule belongs to water shell n if and only if its oxygen atom is within a distance ξ from any of the atoms of protein n. The thickness ξ of the water shells is an input parameter, but not a fitting parameter, of our method. From Eq. (9), we obtain

$$\langle |\tilde{A}(\mathbf{q})|^2 \rangle = \int d\mathbf{r} d\mathbf{r}' e^{-i\mathbf{q}\cdot(\mathbf{r}-\mathbf{r}')} \Bigg\{ \langle A_0(\mathbf{r}) A_0(\mathbf{r}') \rangle + \sum_{n=1}^{N} \sum_{k=1}^{N} \langle A_n(\mathbf{r}) A_k(\mathbf{r}') \rangle + \sum_{n=1}^{N} [\langle A_n(\mathbf{r}) A_0(\mathbf{r}') \rangle + \langle A_0(\mathbf{r}) A_n(\mathbf{r}') \rangle] \Bigg\}. \tag{10}$$

The separation of the electron density in Eq. (9) assumes that the water shells do not overlap, an assumption that may be questionable at high concentrations. We further discuss the situation at high concentrations in Sec. IV A.

In system B, we place the images of the same protein coordinates and separate the instantaneous electron density $B(\mathbf{r})$ into N water droplets and the rest (Fig. 1):

$$B(\mathbf{r}) = \sum_{n=1}^{N} B_n(\mathbf{r}) + B_0(\mathbf{r}). \tag{11}$$

A water molecule belongs to water droplet n if and only if its oxygen atom is within a distance ξ from any of the atoms of the image of protein n. Under this separation, we obtain

$$\langle |\tilde{B}(\mathbf{q})|^2 \rangle = \int d\mathbf{r} d\mathbf{r}' e^{-i\mathbf{q}\cdot(\mathbf{r}-\mathbf{r}')} \Bigg\{ \langle B_0(\mathbf{r}) B_0(\mathbf{r}') \rangle + \sum_{n=1}^{N} \sum_{k=1}^{N} \langle B_n(\mathbf{r}) B_k(\mathbf{r}') \rangle + \sum_{n=1}^{N} [\langle B_n(\mathbf{r}) B_0(\mathbf{r}') \rangle + \langle B_0(\mathbf{r}) B_n(\mathbf{r}') \rangle] \Bigg\}. \tag{12}$$

Now we make a key assumption: the thickness ξ of the water shells is chosen big enough that all the water outside and near the shell boundaries is bulklike. This assumption implies

$$\langle A_0(\mathbf{r}) \rangle = \langle B_0(\mathbf{r}) \rangle \tag{13}$$

and

$$\langle A_0(\mathbf{r}) A_0(\mathbf{r}') \rangle = \langle B_0(\mathbf{r}) B_0(\mathbf{r}') \rangle. \tag{14}$$

The cross term between A_n and A_0 can be written as

$$\langle A_n(\mathbf{r}) A_0(\mathbf{r}') \rangle = \langle A_n(\mathbf{r}) \rangle \langle A_0(\mathbf{r}') \rangle + \alpha_n(\mathbf{r}, \mathbf{r}'), \tag{15}$$

where $\alpha_n(\mathbf{r}, \mathbf{r}')$ is the electron-density correlation between a point $\mathbf{r}$ inside water shell n and $\mathbf{r}'$ outside. Similarly, we write the cross term between B_n and B_0 as

$$\langle B_n(\mathbf{r}) B_0(\mathbf{r}') \rangle = \langle B_n(\mathbf{r}) \rangle \langle B_0(\mathbf{r}') \rangle + \beta_n(\mathbf{r}, \mathbf{r}'), \tag{16}$$

where $\beta_n(\mathbf{r}, \mathbf{r}')$ is the electron-density correlation between a point $\mathbf{r}$ inside water droplet n and $\mathbf{r}'$ outside. The correlation functions $\alpha_n(\mathbf{r}, \mathbf{r}')$ and $\beta_n(\mathbf{r}, \mathbf{r}')$ are significant only near the boundary of the water shell or the water droplet because the correlations diminish when $\mathbf{r}$ and $\mathbf{r}'$ are separated more than the correlation length of liquid water. With ξ sufficiently big, water near the boundaries should be bulklike, implying

$$\alpha_n(\mathbf{r}, \mathbf{r}') = \beta_n(\mathbf{r}, \mathbf{r}'). \tag{17}$$

To rephrase, our key assumption is that the thickness ξ is chosen big enough that Eqs. (13), (14), and (17) are satisfied.

Using Eqs. (13), (14), and (17) and the expansions in Eqs. (10) and (12), we obtain

$$D(\mathbf{q}) = \sum_{n=1}^{N} \sum_{k=1}^{N} [\langle \tilde{A}_n(\mathbf{q}) \tilde{A}_k^*(\mathbf{q}) \rangle - \langle \tilde{B}_n(\mathbf{q}) \tilde{B}_k^*(\mathbf{q}) \rangle] + \sum_{n=1}^{N} [\langle \tilde{A}_n(\mathbf{q}) \rangle \langle \tilde{B}_0^*(\mathbf{q}) \rangle + \langle \tilde{B}_0(\mathbf{q}) \rangle \langle \tilde{A}_n^*(\mathbf{q}) \rangle - \langle \tilde{B}_n(\mathbf{q}) \rangle \langle \tilde{B}_0^*(\mathbf{q}) \rangle - \langle \tilde{B}_0(\mathbf{q}) \rangle \langle \tilde{B}_n^*(\mathbf{q}) \rangle], \tag{18}$$

where the asterisks denote complex conjugates. From Eq. (11), $\langle \tilde{B}_0(\mathbf{q}) \rangle$ can be written as

$$\langle \tilde{B}_0(\mathbf{q})\rangle = \langle \tilde{B}(\mathbf{q})\rangle - \sum_{k=1}^{N} \langle \tilde{B}_k(\mathbf{q})\rangle. \tag{19}$$

Here, $\langle \tilde{B}(\mathbf{q})\rangle = \int d\mathbf{r} e^{-i\mathbf{q}\cdot\mathbf{r}} \langle B(\mathbf{r})\rangle$ is proportional to the Fourier transform of the shape of the entire scattering volume. But the length scale of the scattering volume is too large to be probed by x-ray. Within the range of $\mathbf{q}$ that is measured in experiments, therefore, we can set $\langle \tilde{B}(\mathbf{q})\rangle = 0$ and obtain

$$\langle \tilde{B}_0(\mathbf{q})\rangle = -\sum_{k=1}^{N} \langle \tilde{B}_k(\mathbf{q})\rangle, \tag{20}$$

which is essentially Babinet's principle. Substituting this into Eq. (18) leads to

$$D(\mathbf{q}) = \sum_{n=1}^{N}\sum_{k=1}^{N} D_{nk}(\mathbf{q}), \tag{21}$$

$$D_{nk}(\mathbf{q}) \equiv \langle \tilde{A}_n(\mathbf{q})\tilde{A}_k^*(\mathbf{q})\rangle - \langle \tilde{B}_n(\mathbf{q})\tilde{B}_k^*(\mathbf{q})\rangle - \langle \tilde{A}_n(\mathbf{q})\rangle\langle \tilde{B}_k^*(\mathbf{q})\rangle - \langle \tilde{B}_n(\mathbf{q})\rangle\langle \tilde{A}_k^*(\mathbf{q})\rangle + 2\langle \tilde{B}_n(\mathbf{q})\rangle\langle \tilde{B}_k^*(\mathbf{q})\rangle. \tag{22}$$

The off-diagonal terms, D_{nk} for $n \neq k$, represent the contributions of protein-protein correlations. Assuming that such contributions are negligible (namely, assuming that the proteins do not come close to each other), we can drop the off-diagonal terms:

$$D(\mathbf{q}) = \sum_{n=1}^{N} D_{nn}(\mathbf{q}). \tag{23}$$

This is a safe assumption at low concentrations, but at high concentrations the effect of interprotein correlations may show up in scattering patterns (see Sec. IV A).

C. Averaging over the protein degrees of freedom

The protein degrees of freedom include translations, rotations, and internal motions. The translational degrees of freedom are trivial as long as we ignore protein-protein correlations; $D_{nn}(\mathbf{q})$ is already invariant under translation of each protein molecule. Internal motions can be incorporated, for instance, by sampling from MD simulations or by using normal mode analysis.[14] In this work, however, we focus on the issue of solvation (whether and how much the atomistic-water approach is superior to the continuum-water approach) and defer the issue of internal motions of proteins to the future. Consequently, we are left only with the rotational degrees of freedom.

Rotating a protein molecule is equivalent to rotating the scattering wave vector $\mathbf{q}$ with respect to the protein coordinates fixed in space. The rotational average, therefore, can be expressed as

$$\Delta I(q) = \frac{1}{4\pi}\int d\Omega_{\mathbf{q}} D(\mathbf{q}) = \frac{1}{4\pi}\int d\Omega_{\mathbf{q}} \sum_{n=1}^{N} D_{nn}(\mathbf{q}), \tag{24}$$

where $\int d\Omega_{\mathbf{q}}$ is an integral over the solid angle of $\mathbf{q}$. After the rotational average, each protein contributes to the excess intensity by exactly the same amount, leading to the central formula of this work:

$$\frac{\Delta I(q)}{N} = \frac{1}{4\pi}\int d\Omega_{\mathbf{q}} D_{11}(\mathbf{q}), \tag{25}$$

where $D_{11}(\mathbf{q})$ from Eq. (22) can be rewritten as

$$D_{11}(\mathbf{q}) = |\langle \tilde{A}_1(\mathbf{q})\rangle - \langle \tilde{B}_1(\mathbf{q})\rangle|^2 + [\langle |\tilde{A}_1(\mathbf{q})|^2\rangle - |\langle \tilde{A}_1(\mathbf{q})\rangle|^2] - [\langle |\tilde{B}_1(\mathbf{q})|^2\rangle - |\langle \tilde{B}_1(\mathbf{q})\rangle|^2]. \tag{26}$$

D. Correspondence to the continuum-water theory

The continuum-water theory of x-ray solution scattering is based on the premise that proteins are surrounded by solvation layers of uniform electron density ρ_s which in turn are surrounded by bulk water of uniform electron density ρ_w.[6,7] Below we show that under this premise our central formula of Eqs. (25) and (26) reduces to the continuum-water theory.

A major implication of continuum water is that the entire space of system A can be divided into three disjoint subspaces: the proteins, the solvation layers, and the bulk water. The electron density A_1, then, can be separated into

$$A_1(\mathbf{r}) = A_p(\mathbf{r}) + \rho_s \Theta_s(\mathbf{r}), \tag{27}$$

where A_p is the electron density of the protein and Θ_s is the indicator function for the solvation layer [$\Theta_s(\mathbf{r}) = 1$ if $\mathbf{r}$ is in the solvation layer; $\Theta_s(\mathbf{r}) = 0$ otherwise]. And B_1 can be written as

$$B_1(\mathbf{r}) = \rho_w[\Theta_p(\mathbf{r}) + \Theta_s(\mathbf{r})], \tag{28}$$

where Θ_p is the indicator function for the protein region from which water is excluded.

In this continuum-water picture, the electron density of water is assumed to be uniform and not fluctuate. In Eq. (26), therefore, the second and third terms vanish, and only the first term survives. Our central formula thus reduces to

$$\frac{\Delta I(q)}{N} = \frac{1}{4\pi}\int d\Omega_{\mathbf{q}} |\tilde{A}_p(\mathbf{q}) - \rho_w \tilde{\Theta}_p(\mathbf{q}) + (\rho_s - \rho_w)\tilde{\Theta}_s(\mathbf{q})|^2, \tag{29}$$

which is the main equation for the continuum-water theory of x-ray solution scattering.

III. COMPUTATIONAL DETAILS AND APPLICATIONS

Using myoglobin (Mb) and lysozyme (Lyz) as test cases, we illustrate the application of our atomistic-water method and compare simulated scattering patterns to those observed by WAXS experiments. We took Protein Data Bank structures 1WLA and 6LYZ for Mb and Lyz, respectively, and calculated excess intensities from MD trajectories. Our focus here is the issue of solvation, deferring the investigation of the effects of internal protein motions to future work. Thus, protein coordinates were fixed in the MD simulations.

A. MD simulations

Calculating an excess-intensity pattern requires two MD simulations, one with a protein molecule in a water box and

another with a pure water box. The simulation boxes need to be big enough to accommodate the water shell or the water droplet, which are determined by the thickness ξ. Because we intended to examine a range of values for ξ, we set up simulation boxes somewhat larger than typical MD simulations such that box boundaries were 15 Å apart from protein molecules, which resulted in a box of $59\times71\times74$ Å^3 for Mb and $60\times61\times74$ Å^3 for Lyz. After 20 ps of equilibration, 100 snapshots were collected from 100 ps runs, which were used for the calculation of $\tilde{A}_1(\mathbf{q})$. For $\tilde{B}_1(\mathbf{q})$, we used a water box of size $74\times74\times74$ Å^3. Again, after 20 ps of equilibration, 100 snapshots were collected from a 100 ps run, which were used for both Mb and Lyz.

We used softwares VMD (Ref. 15) and NAMD (Ref. 16) for molecular modeling and simulations, with the TIP3P water model[17] and the CHARMM22 force field.[18] All the MD simulations were done at constant temperature (4 °C, the same as the experimental temperature) and pressure (1 atm). Long-range Coulomb forces were treated with the particle-mesh Ewald method,[19] and 2 fs MD time step was used. Protein coordinates were fixed in all the simulations.

B. Evaluation of ensemble and rotational averages

Calculating the excess intensity $\Delta I(q)$ amounts to evaluating the ensemble averages in Eq. (26) and then the rotational average in Eq. (25). For the evaluation of the ensemble averages involving $\tilde{A}_1(\mathbf{q})$, we use the snapshots from the MD simulation of the protein in water. From each snapshot, $\tilde{A}_1(\mathbf{q})$ is computed by

$$\tilde{A}_1(\mathbf{q}) = \sum_l e^{-i\mathbf{q}\cdot\mathbf{r}_l} f_l(\mathbf{q}), \tag{30}$$

where $\mathbf{r}_l$ is the coordinate of the lth atom contained in the protein or the water shell of thickness ξ. For the atomic form factors $f_l(\mathbf{q})$, we refer to Table VI.1.1.4 in Ref. 20. The form factors therein were obtained under the premise of independent atoms, ignoring the electron-withdrawing effects due to the chemical bonds and electrostatic interactions. In Sec. IV B, we discuss possible pitfalls of using these independent-atom form factors. Let $\tilde{A}_1^{(s)}(\mathbf{q})$ denote the outcome from the sth snapshot.

From the MD simulation of pure water, we obtain $\tilde{B}_1(\mathbf{q})$ through the same summation as in Eq. (30) over the atoms contained in the water droplet. Let $\tilde{B}_1^{(s)}(\mathbf{q})$ denote the outcome from the sth snapshot. We then estimate $D_{11}(\mathbf{q})$ using an unbiased estimator:

$$D_{11}(\mathbf{q}) = |a(\mathbf{q}) - b(\mathbf{q})|^2 + \frac{1}{S}\sum_{s=1}^{S} |\tilde{A}_1^{(s)}(\mathbf{q}) - a(\mathbf{q})|^2 - \frac{S'+1}{S'(S'-1)}\sum_{s=1}^{S'} |\tilde{B}_1^{(s)}(\mathbf{q}) - b(\mathbf{q})|^2, \tag{31}$$

$$a(\mathbf{q}) := \frac{1}{S}\sum_{s=1}^{S} \tilde{A}_1^{(s)}(\mathbf{q}), \quad b(\mathbf{q}) := \frac{1}{S'}\sum_{s=1}^{S'} \tilde{B}_1^{(s)}(\mathbf{q}), \tag{32}$$

where S and S' are the total numbers of snapshots collected from the first and second MD simulations, respectively. The use of the unbiased estimator [namely, the presence of the factor $(S'+1)/(S'-1)$ in Eq. (31)] is more important when a smaller number of snapshots are used.[21]

Evaluating the rotational average in Eq. (25) is a task of spherical quadrature. Among many methods suggested for spherical quadrature, here we use the spiral method,[22] which was also successful in modeling electron spin resonance spectra.[23] A set of J points equally spaced along a spiral on the unit sphere is prepared as

$$\theta_j = \arccos\frac{2j-1-J}{J}, \quad \phi_j = \sqrt{\pi J}\arcsin\frac{2j-1-J}{J}, \tag{33}$$

for $j=1,\ldots,J$. The rotational average is then calculated by taking an average of $D_{11}(\mathbf{q})$ over the solid angles specified by these points. In this work we use $J=1500$, which we find is sufficient for performing the spherical quadrature with less than 1% error up to the resolution of ~2 Å.

Computations of excess-intensity patterns were done in MATLAB. For $\xi=7$ Å, processing of each protein (with 100 snapshots of the protein in a water box and 100 snapshots of a pure water box) took about 30 h on a 2 GHz AMD Opteron processor.

C. Thickness of the water shell

Our method has one parameter to be determined: ξ, the thickness of the water shell. In principle, one could use a water shell of any size as long as it is large enough to contain all the non-bulk-like water. But, in practice, using a thicker water shell means higher computational cost. The best tactic would be to use the smallest value of ξ for which the excess intensity has converged. We examined a range of values for ξ, from 3 to 12 Å, as shown in Fig. 2. The excess intensity $\Delta I(q)$ does converge as ξ increases. The convergence is faster at large q values, which suggests that incorrect representations of solvation shells are likely to affect low scattering angles more than high ones. Overall, we find that $\xi=7$ Å is a reasonable choice for both Mb and Lyz. We expect that 7 Å will be adequate for most proteins, although we cannot rule out the possibility that some proteins may require thicker shells.

D. Comparison to experimental data

With the choice of $\xi=7$ Å, we compare simulated excess-intensity patterns to experimental WAXS data (Figs. 3 and 4). The experiments were performed with 27 mg/ml solution of Mb and 25 mg/ml solution of Lyz at 4 °C. For each protein, scattering intensities were measured seven times from the solution and four times from the pure buffer, from which the excess intensities and error bars were estimated.

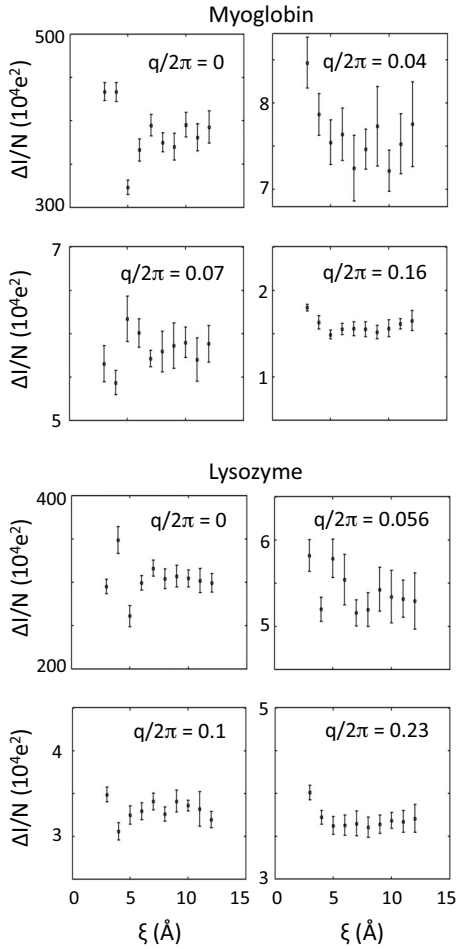

FIG. 2. Choice of ξ, the thickness of the water shell. Excess intensities at various q values are plotted against ξ. Based on these results, we decided on ξ=7 Å.

For more details of the experimental protocols, see Refs. 11 and 13. The error bars of the simulation results were obtained by block averaging (grouping 100 snapshots into 10 blocks).

An excess-intensity pattern typically features a rapidly declining shoulder at low angles and a series of peaks and troughs at high angles. Thus, it is beneficial to plot low- and high-angle regions separately as in Figs. 3 and 4. Logarithmic plots are useful for low angles, but linear plots are more sensible for high angles, not only because excess intensities tend to be negative around $q/2\pi$=0.3/Å but also because logarithmic plots might obscure some of the features at high angles. Scattering angles lower than $q/2\pi$=0.01/Å were not measured in experiments because of the beam stop.

For both Mb and Lyz, we see excellent agreements between the simulated and observed data, except for some discrepancies beyond $q/2\pi$=0.2/Å. This level of agreement in the range of $0.01/\text{Å} \leq q/2\pi \leq 0.2/\text{Å}$, corresponding to the length scale between 5 and 100 Å, is unprecedented and indicates that our atomistic-water method has correctly captured the nature of solvation around proteins that the previous continuum-water methods have missed. In Sec. IV B, we explore possible sources of the residual discrepancies.

For comparison, in Figs. 3 and 4, we also show the scattering patterns produced by CRYSOL,[7] which is based on the continuum description of water. As explained in Sec. II A, one may argue that the scattering patterns obtained with continuum-water models are compatible with the scaled background-subtraction scheme [Eq. (6)] rather than the direct scheme [Eq. (5)]. It is, therefore, somewhat ambiguous to compare our atomistic-water patterns and the CRYSOL patterns at high angles ($q/2\pi \geq 0.2/\text{Å}$) where the two schemes yield significantly different results. (See Ref. 11 for examples of comparing CRYSOL patterns with experimental data using the scaled scheme.) Nevertheless, the following observations are valid regardless of the issue of background subtraction. (1) The atomistic-water calculations are much more accurate in reproducing peaks and troughs of the experimental scattering patterns. (2) The CRYSOL patterns tend to "overshoot" toward the zero angle. This overshooting is clearly absent in the atomistic-water patterns.

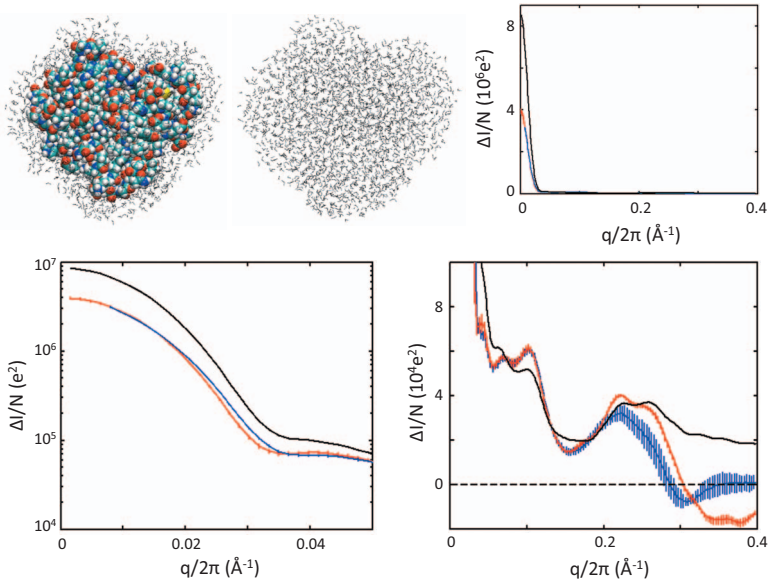

FIG. 3. (Color) Simulated and observed excess-intensity patterns for myoglobin. Top left: the protein plus a water shell of thickness 7 Å. Top middle: a corresponding water droplet. Top right: the excess-intensity patterns for $0 \leq q/2\pi \leq 0.4/\text{Å}$. Bottom left: a small-angle region in logarithmic scale. Bottom right: a wide-angle region in linear scale. Blue is the experimental data, red is our computational result, and black is the pattern produced by CRYSOL.

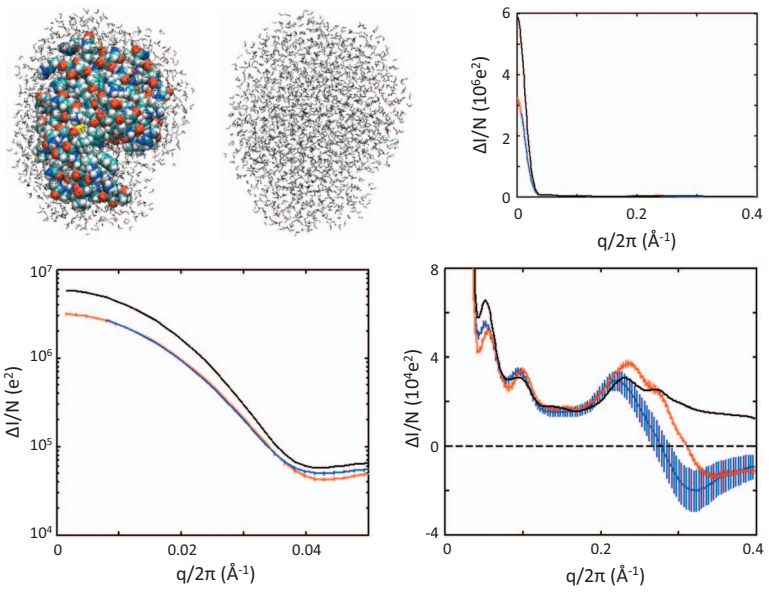

FIG. 4. (Color) Simulated and observed excess-intensity patterns for lysozyme. Top left: the protein plus a water shell of thickness 7 Å. Top middle: a corresponding water droplet. Top right: the excess-intensity patterns for $0 \leq q/2\pi \leq 0.4/\text{Å}$. Bottom left: a small-angle region in logarithmic scale. Bottom right: a wide-angle region in linear scale. Blue is the experimental data, red is our computational result, and black is the pattern produced by CRYSOL.

IV. DISCUSSION

A. High concentrations

Our atomistic-water theory is based on two assumptions: (1) nonoverlapping water shells of an identical thickness can be placed around proteins such that all the water outside and near the shell boundaries can be considered bulklike, and (2) proteins are well separated from each other so that protein-protein correlations have negligible impacts on scattering patterns. Both assumptions are reasonable at dilute conditions, and simulated scattering patterns are expected to agree best with those observed at low concentrations. Performing experiments at high concentrations do, however, offer benefits. Because the magnitude of the excess intensity [Eq. (25)] is proportional to the number of proteins in the scattering volume, high concentrations yield better signal-to-noise ratios. High concentrations are also of interest pertaining to internal motions of proteins.[13] Understanding how excess intensities may change at high concentrations is, therefore, important.

The effect of high concentrations can be threefold. First, interprotein correlations may alter the excess intensity. The nature of the alteration depends on the characteristics of the protein-protein interactions: whether they are attractive or repulsive, whether they depend on the relative orientation of proteins, and so on. The most prevalent characteristic is the hard-core repulsion (proteins cannot go through each other), which tends to decrease scattering intensities at low angles.[2] Second, crowding may change the nature of solvation layers around proteins. Such changes can have substantial impacts on scattering patterns, as Fig. 2 suggests. Third, crowding may also affect intraprotein correlations by suppressing internal protein motions; proteins of higher flexibility should be more susceptible to such effects. The interprotein effects are expected to show up only at low angles (namely, at length scales larger than the protein size), but the solvation layer and the intraprotein effects may appear at higher angles.

B. Possible sources of the residual discrepancies

The remaining discrepancies between simulated and observed excess-intensity patterns (Figs. 3 and 4) may be attributed to a few possible sources.

(1) The atomic form factors, obtained under the premise of independent atoms, may not be adequate because they do not account for the electron-withdrawing effects. This issue appears especially important for water; refinement of atomic form factors has been shown to significantly improve the accuracy of simulated scattering patterns of liquid water.[24–26] We anticipate that a large portion of the discrepancies at high angles ($q/2\pi \geq 0.2/\text{Å}$) will be removed by improving the form factors for water.

(2) The water model, TIP3P in the present case, may not be perfect in representing the atomic coordinates of water. In fact, the use of more sophisticated water models (such as TIP5P) appears to yield better agreements with x-ray scattering data.[25,26] This issue is, however, somewhat difficult to deal with because most of the current protein force fields are designed to be used with the TIP3P water model.

(3) In this work, to focus on the issue of solvation, we fixed protein coordinates in all the calculations. The simulated scattering patterns thus do not capture the effects of internal protein motions. Such effects, we expect, will be more important for flexible proteins than rigid ones. We intend to address this issue by using various methods such as B-factor analysis, MD, and normal mode analysis.

(4) The high-concentration effects we discussed above may also play a role. But, we suspect that those effects should be minor at the conditions of the experimental data used here (27 mg/ml for Mb and 25 mg/ml for Lyz).

V. CONCLUSION

Although the full information content of solution-scattering data is yet to be determined, clearly there exists vastly more information in the WAXS region than in the SAXS region, as evidenced by comparing the scattering patterns of Mb (Fig. 3) and Lyz (Fig. 4). Previous computational methods, however, have been unable to calculate accurate scattering patterns in the WAXS regime. The atomistic-water method presented here can produce scattering patterns of high accuracy throughout the regions of SAXS and WAXS. We expect that this method will be a basis for further development of computational methods, toward making solution scattering a powerful tool for the study of protein structures.

ACKNOWLEDGMENTS

We thank Mihai Anitescu and Sichun Yang for many insightful discussions. This research was supported by the U.S. Department of Energy, under Contract No. DE-AC02-06CH11357.

[1] S. Doniach, Chem. Rev. (Washington, D.C.) **101**, 1763 (2001).
[2] C. D. Putnam, M. Hammel, G. L. Hura, and J. A. Tainer, Q. Rev. Biophys. **40**, 191 (2007).
[3] D. M. Tiede, R. Zhang, and S. Seifert, Biochemistry **41**, 6605 (2002).
[4] R. F. Fischetti, D. J. Rodi, D. B. Gore, and L. Makowski, Chem. Biol. **11**, 1431 (2004).
[5] L. Makowski, D. J. Rodi, S. Mandava, S. Devarapalli, and R. F. Fischetti, J. Mol. Biol. **383**, 731 (2008).
[6] R. D. B. Fraser, T. P. MacRae, and E. Suzuki, J. Appl. Crystallogr. **11**, 693 (1978).
[7] D. Svergun, C. Barberato, and M. H. J. Koch, J. Appl. Crystallogr. **28**, 768 (1995).
[8] E. Lattman, Proteins **5**, 149 (1989).
[9] F. Merzel and J. C. Smith, Acta Crystallogr., Sect. D: Biol. Crystallogr. **58**, 242 (2002).
[10] T. Head-Gordon and G. Hura, Chem. Rev. (Washington, D.C.) **102**, 2651 (2002).
[11] R. F. Fischetti, D. J. Rodi, A. Mirza, T. C. Irving, E. Kondrashkina, and L. Makowski, J. Synchrotron Radiat. **10**, 398 (2003).
[12] P. M. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics* (Cambridge University Press, Cambridge, England, 1995).
[13] L. Makowski, D. J. Rodi, S. Mandava, D. D. L. Minh, D. B. Gore, and R. F. Fischetti, J. Mol. Biol. **375**, 529 (2008).
[14] *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*, edited by Q. Cui and I. Bahar (CRC, Boca Raton, FL, 2006).
[15] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[16] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, J. Comput. Chem. **26**, 1781 (2005).
[17] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[18] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiýýrkiewicz-Kuczera, D. Yin, and M. Karplus, J. Phys. Chem. B **102**, 3586 (1998).
[19] T. Darden, D. York, and L. Pedersen, J. Chem. Phys. **98**, 10089 (1993).
[20] *International Tables for Crystallography*, edited by A. J. C. Wilson (Kluwer Academic, Boston, 1992), Vol. C.
[21] J. F. Kenney and E. S. Keeping, *Mathematics of Statistics*, 2nd ed. (Van Nostrand, Princeton, 1951).
[22] A. Ponti, J. Magn. Reson. **138**, 288 (1999).
[23] D. Sezer, J. H. Freed, and B. Roux, J. Chem. Phys. **128**, 165106 (2008).
[24] K. Hermansson, Chem. Phys. Lett. **260**, 229 (1996).
[25] J. M. Sorenson, G. Hura, R. M. Glaeser, and T. Head-Gordon, J. Chem. Phys. **113**, 9149 (2000).
[26] M. Krack, A. Gambirasio, and M. Parrinello, J. Chem. Phys. **117**, 9409 (2002).