Modeling solution X-ray scattering of biomacromolecules using an explicit solvent model and the fast Fourier transform

Tong, Dudu; Yang, Jianbin; Lu, Lanyuan

2015

Tong, D., Yang, J., & Lu, L. (2015). Modeling solution X-ray scattering of biomacromolecules using an explicit solvent model and the fast Fourier transform. *Journal of Applied Crystallography*, 48, 1834-1842.

<https://hdl.handle.net/10356/81221>

<https://doi.org/10.1107/S1600576715018816>

© 2015 International Union of Crystallography. This paper was published in *Journal of Applied Crystallography* and is made available as an electronic reprint (preprint) with permission of International Union of Crystallography. The published version is available at: [<http://dx.doi.org/10.1107/S1600576715018816>]. One print or electronic copy may be made for personal use only. Systematic or multiple reproduction, distribution to multiple locations via electronic or other means, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper is prohibited and is subject to penalties under law.

Downloaded on 26 Apr 2023 21:58:28 SGT

Modeling solution X-ray scattering of biomacromolecules using an explicit solvent model and the fast Fourier transform

Dudu Tong,^a Jianbin Yang^b and Lanyuan Lu^{a*}

Received 31 May 2015
Accepted 6 October 2015

Edited by D. I. Svergun, European Molecular Biology Laboratory, Hamburg, Germany

Keywords: X-ray solution scattering; small-angle X-ray scattering (SAXS); wide-angle X-ray scattering (WAXS); proteins; structural dynamics.

Supporting information: this article has supporting information at journals.iucr.org/j

^aSchool of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore, 637551, Singapore, and ^bCollege of Sciences, Hohai University, Fucheng West Road 8, Jiangning, Nanjing, Jiangsu Province 211100, People's Republic of China. *Correspondence e-mail: lylu@ntu.edu.sg

A novel computational method based on atomic form factors and the fast Fourier transform (FFT) is developed to compute small- and near-wide-angle X-ray scattering profiles of biomacromolecules from explicit solvent modeling. The method is validated by comparing the results with those from non-FFT approaches and experiments, and good agreement with experimental data is observed for both small and near-wide angles. In terms of computational efficiency, the FFT-based method is advantageous for protein solution systems of more than 3000 atoms. Furthermore, the computational cost remains nearly constant for a wide range of system sizes. The FFT-based approach can potentially handle much larger molecular systems compared with popular existing methods.

1. Introduction

Solution X-ray scattering, including small-angle X-ray scattering (SAXS) and wide-angle X-ray scattering (WAXS), is a widely used experimental method to probe the conformations of biomacromolecules (Koch *et al.*, 2003; Svergun & Koch, 2003; Putnam *et al.*, 2007; Mertens & Svergun, 2010; Graewert & Svergun, 2013; Blanchet & Svergun, 2013; Lipfert & Doniach, 2007). SAXS/WAXS has a variety of applications in structural biology, such as understanding protein conformational changes (Fischetti *et al.*, 2004; Tainer, 2011; Makowski, Gore *et al.*, 2011; Makowski, Bardhan *et al.*, 2011; Oroguchi *et al.*, 2009; Zheng & Tekpinar, 2011). Traditionally, SAXS experiments can only reach values of scattering vector magnitude q up to 5 nm^{-1} , containing information on radius of gyration, on the approximate shape of the solute molecule and, rarely, on some secondary structures (Koch *et al.*, 2003). In recent years, advances in the new generation of light source and detection devices have made it possible to measure data in the wide-angle region, providing higher-resolution structure details (Graewert & Svergun, 2013).

The outcome of a SAXS/WAXS experiment is a one-dimensional intensity curve, which typically contains insufficient information to solve the three-dimensional solute structure. Thus, computational tools are needed to extract structure details from SAXS/WAXS profiles. There are already many methods in the literature that evaluate SAXS/WAXS profiles from structure models (Svergun *et al.*, 1995; Schneidman-Duhovny *et al.*, 2010; Konarev *et al.*, 2006; Merzel & Smith, 2002; Bardhan *et al.*, 2009; Oroguchi *et al.*, 2009; Park *et al.*, 2009; Yang *et al.*, 2009; Grishaev *et al.*, 2010; Stovgaard *et al.*, 2010; Poitevin *et al.*, 2011; Liu *et al.*, 2012; Ravikumar *et al.*, 2013; Chen & Hub, 2014; Niebling *et al.*, 2014; Olds &

Duxbury, 2014; Kofinger & Hummer, 2013; Schneidman-Duhovny *et al.*, 2013; Nguyen *et al.*, 2014; Chen & Hub, 2015). These methods mainly differ in two aspects: in the treatment of contributions from the hydration layer and the displaced solvent; and in the method of orientational averaging to obtain the one-dimensional scattering profile. For example, *Crysol* (Svergun *et al.*, 1995) implicitly models the hydration shell with a layer of uniformly distributed excess electron density. The displaced solvent is also modeled implicitly using the excluded volume of all solute atoms, and the orientational averaging is performed using the spherical harmonic expansion. Similar to *Crysol*, a number of other approaches adopt various implicit solvent treatments (Schneidman-Duhovny *et al.*, 2010; Konarev *et al.*, 2006; Merzel & Smith, 2002; Bardhan *et al.*, 2009; Poitevin *et al.*, 2011; Liu *et al.*, 2012). As the atomistic structure and distribution of water should affect the high- q (wide-angle) region, these methods usually target q values up to 5 nm^{-1} .

Conversely, *Fast-SAXS-pro* (Ravikumar *et al.*, 2013) implements an explicit representation of the solvent molecules in the hydration layer, while the displaced solvent is still implicitly modeled using excluded volumes. As a further improvement, the *AXES* method (Grishaev *et al.*, 2010) models both the hydration shell and the excluded solvent with explicit water molecules and achieves good agreement with experiment for q up to 10 nm^{-1} (near-WAXS region). However, in both approaches the positions of the water molecules in the hydration layer are actually taken from a pure water simulation, neglecting the structure difference between the bulk and the hydration layer.

To include the explicit solvent molecules in a different way from *AXES*, there are other methods using molecular dynamics (MD) simulation trajectories of both protein solution and pure solvent. These methods can correctly represent the structure of the solvent molecules in the hydration layer, and the water density distribution and fluctuation are also considered in the calculations. In general, these methods often have a good agreement with experimental data up to 10 nm^{-1} . However, these methods require MD simulations of proteins in solvent, which costs additional computational time. For example, Park *et al.* (2009) proposed a formulation to compute scattering intensities from MD simulations. The computational time for computing scattering profiles using this method scales with the number of atoms N , which is superior to the Debye formula that scales with N^2 . However, for large N the computation can still be expensive. Later, Chen & Hub (2014) used the same method to build a web server called *WAXSiS*, which takes protein conformational fluctuations into consideration. The results from both Park *et al.* and Chen *et al.* are consistent with experimental near-WAXS profiles up to 10 nm^{-1} . Additionally, Oroguchi *et al.* (2009) proposed a similar formulation, but spherical harmonics functions are adopted for the calculations on scattering profiles. Very recently, Kofinger & Hummer (2013) developed an explicit solvent approach using a histogram-based Debye method for modeling near-WAXS and WAXS profiles. This method still has $O(N^2)$ computational complexity when calculating the

distance histograms. But, as there is no need to evaluate sine functions in the histogram-calculation step, the prefactor of $O(N^2)$ is small and the method is very efficient for certain system sizes.

In this paper, we propose a new method based on the three-dimensional fast Fourier transform (FFT) to calculate the scattering intensity profiles of biomacromolecules. In the literature, FFT has been implemented in SAXS/WAXS calculations, and typically the transform starts from a three-dimensional density map (*e.g.* Schmidt-Rohr, 2007). However, in this article we develop an efficient way to directly incorporate the atomic coordinates and form factors, without first explicitly converting them to a density map. Our approach is particularly convenient for biological systems as the atomic coordinates are usually available from Protein Data Bank and MD simulations, while it avoids the approximation for converting atomic coordinators to density values (Rey & Dumas, 1984). The FFT-based method is implemented with two explicit solvent approaches, with and without protein MD simulations, and the results are compared with those from the *WAXSiS* and *AXES* servers, which are non-FFT based. For instance, we follow the treatment of displaced and hydrated solvent molecules proposed by Park *et al.* (2009) and use FFT to conduct Fourier transformation and orientational averaging, thus improving the computational efficiency of the MD-based approach. The foremost advantage of our method is its high computational efficiency.

2. Theory and computational details

2.1. Envelope construction

In SAXS/WAXS experiments, the scattering profile is usually measured for two systems: the protein solution system and the pure solvent system. The final scattering profile is obtained by computing the difference of these two. In computer modeling, if A and B represent the protein solution system and the pure solvent system, respectively, the resulting scattering profile $I(q)$ can be calculated from the Fourier transforms of the electron densities of two ‘envelopes’ $A(\mathbf{q})$ and $B(\mathbf{q})$ (Park *et al.*, 2009; Chen & Hub, 2014). The envelope for the protein solution A includes the solute and a layer of solvent molecules, while the solvent envelope of B created from a separate solvent simulation has the same shape as that of A . Here, a spherical envelope is used for both systems for convenience. The center of the sphere is placed at the center of the maximum diameter of the protein molecule. The radius $r = D_{\text{max}}/2 + d_{\text{shell}}$, where D_{max} is the maximum diameter of the protein molecule and d_{shell} is the thickness of the solvent shell. For the pure solvent system, the spherical envelope is constructed using the same center and radius. In previous work (Park *et al.*, 2009; Chen & Hub, 2014), the envelope of A is constructed by including the protein molecule and water molecules within d_{shell} to the surface of protein, while an image of a protein with identical atomic coordinates is placed in system B and the envelope of B contains all water molecules included within d_{shell} to the surface of the image of the protein.

To have a fair comparison of the computational efficiency with the methods reported in the literature, we have also implemented the latter type of envelope for the results in §3.2.

Following the derivations from earlier work (Park *et al.*, 2009; Chen & Hub, 2014) we have

$$I(q) = \langle D(\mathbf{q}) \rangle_{\Omega} \quad (1)$$

and

$$D(\mathbf{q}) = \left\langle |A(\mathbf{q})|^2 \right\rangle_{\omega} - \left\langle |B(\mathbf{q})|^2 \right\rangle_{\omega} - 2\text{Re}[\langle B^*(\mathbf{q}) \rangle_{\omega} \langle A(\mathbf{q}) - B(\mathbf{q}) \rangle_{\omega}]. \quad (2)$$

Here, the subscript Ω denotes the ensemble averages over the solute orientational degrees of freedom, while ω refers to the solute conformational degrees of freedom and the solvent degrees of freedom. The third term of equation (2) is the real part of a complex number, where the symbol '*' denotes the complex conjugate. Although equation (2) requires the information only within the two envelopes, $I(q)$ is actually computed to represent the intensity difference between the two molecular systems with infinite sizes.

The Fourier transform of the electron density inside an envelope can be calculated by the following formula:

$$A(\mathbf{q}) = \sum_j^N f_j(q) \exp(-i\mathbf{q} \cdot \mathbf{r}_j), \quad (3)$$

where j is the index of an atom inside the envelope, N is the total number of atoms inside the envelope, $\mathbf{r}_j$ is the coordinate of the j th atom and $f_j(q)$ is the atomic form factor (Cromer & Waber, 1965; Sorenson *et al.*, 2000) of the j th atom at the scattering vector magnitude q .

2.2. FFT-based orientational averaging

Instead of using the spherical average method (Chen & Hub, 2014; Park *et al.*, 2009), we adopt an FFT-based approach to evaluate the Fourier transforms $A(\mathbf{q})$, $B(\mathbf{q})$ and the final one-dimensional scattering intensity.

From equation (3), the FFT cannot be directly applied as the right-side term of the equation contains $f_j(q)$, which is related to the atom type of atom j and the scalar q . So the system is divided into several subsystems, each containing one type of atom. For each subsystem, we first evaluate $A^t(\mathbf{q})$, which is defined as

$$A^t(\mathbf{q}) = \sum_j^{N_t} \exp(-i\mathbf{q} \cdot \mathbf{r}_j), \quad (4)$$

where t is the index of the subsystem and N_t is the number of atoms in this subsystem. $A(\mathbf{q})$ can be obtained by the summation of $A^t(\mathbf{q})$, as shown in the following equation:

$$A(\mathbf{q}) = \sum_t A^t(\mathbf{q}) f^t(q). \quad (5)$$

$A^t(\mathbf{q})$ can be discretized and evaluated by the FFT algorithm. Cubic grids are built in both real and reciprocal space. The bin width and maximum value of r and q satisfy the relationship that $r_{\text{bin}} = 2\pi/(2q_{\text{max}})$ and $q_{\text{bin}} = 2\pi/r_{\text{max}}$. This allows us to rewrite equation (4) in the form of a discrete Fourier trans-

form, to which the FFT algorithm can be applied. The atom positions $\mathbf{r}_j$ are then discretized to nearby grid points and the particle density of each grid is calculated. Equation (4) is then rewritten as

$$A^t(h, k, l) = \sum_a \sum_b \sum_c \left\{ \rho(a, b, c) \times \exp[-2\pi i(h, k, l) \cdot (a, b, c)/N_g^3] \right\}. \quad (6)$$

Here, the vector (a, b, c) contains the indices of grid points in the real space, and (h, k, l) consists of the corresponding indices in the reciprocal space. ρ is the number of type t atoms at grid point (a, b, c) , and N_g is the number of grid points in each dimension. We then apply the FFT algorithm to evaluate A^t of all grid points at once.

A problem here is that the simulation box is usually not very large, at the length scale of several nanometres. According to the relationship mentioned above, q_{bin} will be quite large (e.g. $r_{\text{max}} = 10$ nm corresponds to $q_{\text{bin}} \simeq 0.6$ nm⁻¹). This results in very low resolution in q and causes difficulty in comparing with experimental data. To solve this problem, we apply a method called zero padding (Castiglioni, 2014), which is widely used in digital signal processing, to increase the resolution of the reciprocal space. To implement the zero padding, we first transform the low-resolution $A(\mathbf{q})$ back to $A(\mathbf{r})$, which is the electron density of the system. Then the original system is expanded to a much larger box for the desired q_{bin} . In this study we use $q_{\text{bin}} = 0.2$ nm⁻¹, corresponding to $r_{\text{max}} \simeq 31.4$ nm. The electron density of the larger box is denoted as $A(\mathbf{R})$, with the expanded volume filled with zero density as no atoms are actually there. Finally, the Fourier transform of $A(\mathbf{R})$ is performed to obtain the scattering intensity with higher resolution. This step is the most time-consuming part in the whole algorithm, because the number of grid points is quite large for the following two reasons. On one hand, q_{bin} must be small enough to ensure that the output data are comparable to the experimental counterpart. On the other hand, r_{bin} must also be small enough to ensure the accuracy of the real-space discretization, which forces us to use a much larger q_{max} than we actually need. Note that increasing the number of atom types will not significantly increase the computational time as $A(\mathbf{R})$ is computed from the sum of all atom types. Typically, in standard experimental WAXS data q_{max} is 15 or 20 nm⁻¹. However, in the FFT approach, q_{max} is usually set to be 30 nm⁻¹ or even larger to ensure the real-space accuracy. Nevertheless, we show in §3 that the computational efficiency of our FFT-based method is still higher than the spherical average method for a wide range of system sizes.

After $A(\mathbf{q})$ and $B(\mathbf{q})$ have been obtained for all the grid points in q space, the experiential intensity $D(\mathbf{q})$ for a particular scattering vector $\mathbf{q}$ can be computed (Chen & Hub, 2014; Park *et al.*, 2009). Then we need to project the three-dimensional scattering intensity onto a one-dimensional scattering profile for the scalar value q . This is accomplished by a method called nearest neighbor searching, in which the scattering intensity of each grid point is added to the output q value nearest to its modulus. The procedure may be improved by

channel sharing (Schmidt-Rohr, 2007). However, we continue to use the nearest neighbor searching method as no significant improvement from channel sharing was observed with our data.

It is noticed that, apart from the evaluation of scattering intensity, the MD simulation also takes a large proportion of the total computational time. However, the FFT-based method can be applied to SAXS/WAXS-guided MD simulations (Chen & Hub, 2015) to reduce the overall computation time, where the scattering intensity is evaluated on the fly.

2.3. FFT with a non-simulation-based method

The superiority of the FFT-based method is more obvious if expensive MD simulations are not required in the calculation of scattering intensities. Thus, we further generalize the FFT-based method with non-simulation-based methods to compute the scattering intensity. Specifically, we implemented the FFT algorithm in the method used in the *AXES* web server. The *AXES* method overlaps the target protein structure with a pre-simulated trajectory of a water box to determine the position of displaced and hydrated water molecules. The scattering intensity is thus calculated using the following formula:

$$I(q) = \left\langle \left\langle \left| A^{\text{mol}}(\mathbf{q}) - A^{\text{disp}}(\mathbf{q}) + \delta\rho A^{\text{surf}}(\mathbf{q}) \right|^2 \right\rangle_{\Omega/\text{solv}} \right\rangle_{\text{ens}}, \quad (7)$$

where $A^{\text{mol}}(\mathbf{q})$, $A^{\text{disp}}(\mathbf{q})$ and $A^{\text{surf}}(\mathbf{q})$ are the Fourier transforms of the electron densities of the protein molecule, displaced water molecules and hydration water molecules, respectively. $\delta\rho$ is the excess electron density of the protein hydration shell compared with bulk water. $I(q)$ is obtained by averaging the resulting intensity over angular orientations Ω , solvent distributions and protein conformations. In the original *AXES* method, the orientational average is performed numerically using the Fibonacci grid method, where 1589 equi-spaced $\mathbf{q}$ directions are used, similar to the spherical average method. As there are explicit water molecules included in the calculation, the computation is also time consuming. As discussed in §2.2, $A(\mathbf{q})$ can be easily calculated by the FFT algorithm, thus increasing the computational efficiency drastically.

2.4. MD simulation details

We have chosen three proteins as our benchmark systems for which WAXS profiles are available in the Small Angle Scattering Biological Data Bank (SASBDB) (Valentini *et al.*, 2015). The initial structures for MD simulations were obtained from the Protein Data Bank (PDB; <http://www.rcsb.org/>). The chosen proteins are lysozyme (PDB code 1lyz; Nguyen *et al.*, 2014), myoglobin (PDB code 1wla; Maurus *et al.*, 1998) and RNase (PDB code 1c0b; Bell, 1999). Missing side chains were completed using the function module on the *WHAT IF* server (Hekkelman *et al.*, 2010). *Gromacs* 4.5.5 (Pronk *et al.*, 2013) and the *CHARMM* (Rey & Dumas, 1984) force field were used to perform all the atomistic simulations with large dodecahedral simulation boxes around the proteins. The nearest distance between the protein molecule and the box

boundary was set to 5 nm, and the resulting box dimensions were about 15 nm. We have chosen such a large box size because we want to study the effect of different envelope size. The protein molecules were solvated with TIP3P water molecules (Jorgensen *et al.*, 1983) and neutralized by sodium or chlorine counter-ions. A short constant NPT simulation of 100 ps was performed to equilibrate the system. The final constant NPT production run was conducted for 1 ns. The simulated conformations were collected every 10 ps, resulting in 100 frames for the scattering calculation of one protein. Additionally, we performed a simulation for pure solvent with a similar box size to calculate $B(\mathbf{q})$. To study the effect of protein conformational change, we also performed shorter protein backbone restraint simulations of 200 ps under a constant NPT condition.

2.5. Density correction

As the bulk water density during the simulation is usually slightly different from the experimental value (334 e nm^{-3}), we need to perform density correction to obtain more accurate scattering intensity curves. We scale the electron density of water inside the envelope by a ratio of $\delta_{A,B} = \rho_{\text{exp}}/\rho_{A,B}^{\text{bulk}}$ for protein solution and pure solvent systems, respectively. Here, ρ_{exp} is the experimental water electron density, and ρ^{bulk} is the water electron density outside the envelope.

2.6. Comparison between scattering profiles

To measure the discrepancy between two scattering profiles from experiments (I_{exp}) and theoretical calculations (I_{cal}), we minimize χ^2 , defined as follows:

$$\chi^2 = N_q^{-1} \sum_{i=1}^{N_q} \left[\frac{I_{\text{exp}}(q_i) - fI_{\text{cal}}(q_i) + c}{\sigma(q_i)} \right]^2. \quad (8)$$

Here, N_q is the number of points in the scattering intensity curve, f is the parameter to scale the computed curve in order to fit the experimental data and c is the parameter to reduce the error from the experimental buffer subtraction process (Chen & Hub, 2014). A smaller χ^2 means better fitting of the two curves. Before minimizing χ^2 , some preprocessing is performed for both experimental and computed scattering intensities. For experimental profiles, data points having negative values or deviating greatly from the average curves are ignored. Usually, the resolution in q of the experimental data is higher than that of the computed profiles, so we perform interpolation on $\log I_{\text{cal}}(q)$ using cubic splines.

3. Results and discussion

3.1. Accuracy of the FFT-based method

Before performing any calculations we first need to fine-tune the FFT-related parameters for optimal performance and acceptable precision. In this part, the result of our FFT-based approach is compared with those from the Debye summation and the spherical average method. The Debye summation is without any approximation and was chosen to calculate the

standard curve. We used protein conformations without solvent to avoid high computational cost. The three benchmark proteins are lysozyme, myoglobin and RNase. Because the results of Debye summation do not have the standard error needed by equation (8), an empirical standard error $\sigma(q_i) = I(q_i)(q_i + \alpha)\beta$ is used to evaluate the χ^2 value between different computed scattering intensities. Here, we choose $\alpha = 0.15$ and $\beta = 0.3$ following Stovgaard *et al.* (2010).

Fig. 1(a) shows the variation of the χ^2 value between the scattering profiles calculated by the Debye summation and the

FFT-based method when the value of $q_{\max}$ changes. It is seen that a larger $q_{\max}$ value leads to a smaller χ^2 value. This is because a larger $q_{\max}$ value corresponds to a smaller r_{bin} , suggested by the relationship $r_{\text{bin}} = 2\pi/(2q_{\max})$. A smaller r_{bin} reduces the error caused by discretizing the atom coordinates. However, on the other side, a larger $q_{\max}$ also leads to more computational cost. In our FFT-based calculations, q_{bin} is usually fixed to be 0.2 nm^{-1} for the necessary resolution in q space. We notice that the χ^2 values become very small (4.496×10^{-2} , 5.235×10^{-2} and 4.686×10^{-2} for the three proteins) at $q_{\max} = 30 \text{ nm}^{-1}$, and the further decrease is not obvious after this $q_{\max}$ value. Additionally, the calculated scattering profiles at the $q_{\max}$ value are plotted in Fig. 1(b) and compared with the profiles from other methods. From Fig. 1(b), the scattering profiles calculated by the spherical average method fit the standard curves perfectly with the χ^2 values 3.068×10^{-5} , 3.433×10^{-5} and 5.528×10^{-5} for the three proteins. Although, the χ^2 values of the FFT-based method are slightly larger, the scattering curves of the FFT-based method overlap very well with the curves from the Debye method, and the difference is indistinguishable within the line width. Thus, $q_{\max}$ is chosen to be 30 nm^{-1} for the following computations in this paper, with $q_{\text{bin}} = 0.2 \text{ nm}^{-1}$.

3.2. Computational efficiency of the FFT-based method

We tested the computational time consumed by our FFT-based method for different systems and compared the results with those from the spherical average method. The calculations were performed on the envelopes usually constructed in scattering modeling, which include proteins and nearby solvent molecules. The 101 frames (including the initial frame) in the production simulation trajectories of the three proteins were used, with the thicknesses of the hydration shells ranging from 0.5 to 3 nm. To compare fairly with the original spherical average method, we constructed protein-shaped envelopes with the same hydration layer thicknesses to perform calculations using both methods. The two algorithms were implemented by using our homemade *Gromacs* analysis tools, and all computations were conducted on a single Intel Xeon CPU (E7-8850 at 2.00 GHz). The computational times are shown in Fig. 2. The results for the three proteins are colored in red (lysozyme), green (myoglobin) and blue (RNase), while spherical points and cubic points represent the spherical average and FFT, respectively. The seven points on each curve represent different thicknesses of solvent shell, ranging from 0.5 to 3.0 nm, corresponding to different total atom numbers shown in the x axis.

It should be noted that a small protein molecule plus a thick hydration shell can be equivalent to a large protein molecule with a thin hydration shell, in terms of computational cost required to evaluate scattering intensities. Because there are no reliable near-WAXS data in SASDB for very large protein molecules, we mimic large molecular systems by simply increasing the thickness of the hydration shell for our benchmark systems. From Fig. 2, the computational time consumed by the spherical average method increases linearly

Figure 1
(a) The χ^2 values between the scattering profiles calculated by the spherical average method and our FFT-based method, using different $q_{\max}$ values. (b) The scattering profiles calculated by the spherical average method, the FFT-based method and the Debye summation.

as the number of atoms increases. However, the computational time remains nearly unchanged as the system size increases for the FFT-based method. This is consistent with our discussion in §2.2. The spherical average method has the computational complexity scale as $O(JNn_q)$, where J is the number of orientations ($J = 1000$ in this calculation). By contrast, the scaling of the FFT-based method is $O[N_{\text{FFT}}^3 \log(N_{\text{FFT}})]$. Here, N_{FFT} is the number of grid points on one dimension for both real and reciprocal spaces after zero padding. For the real space, $N_{\text{FFT}} = r/r_{\text{bin}}$, where r is just 31.4 nm if the envelope dimension does not exceed r_{max} . Note that r_{bin} is fixed for the necessary real space precision ($r_{\text{bin}} \approx 0.105$ nm for $q_{\text{max}} = 30$ nm⁻¹). Thus, when the range of output q is fixed, the computational time for the spherical average method linearly depends on the number of atoms N , while the computational time for the FFT-based method remains a constant with the change in number of atoms. From Fig. 2, it is also estimated that the computational efficiencies of the two methods will be approximately the same for systems of around 3000 atoms and with further increase in atom number the computational efficiency of FFT becomes better. For instance, lysozyme has 1960 protein atoms, and the total number of envelope atoms is around 4444 after including a 0.5 nm water layer, thus the FFT-based method requires less computational time for this case. In general, the FFT-based method becomes superior for most protein plus solvent systems.

The size-independent computational efficiency shown in Fig. 2 applies to molecular systems (proteins plus hydration shells) with diameters smaller than $r_{\text{max}} = 31.4$ nm, which is a result of zero padding for $q_{\text{bin}} = 0.2$ nm⁻¹. This corresponds to a protein molecular weight of about 10⁴ kDa, if we set the thickness of the hydration shell to be 0.7 nm for a spherical envelope and assume an average protein density of 1.22 g cm⁻³. Thus, our r_{max} should be large enough for most

Figure 2

The computational time required by the spherical average method and our FFT-based method for different hydration shell thicknesses. The seven points on each curve represent 0.5, 0.7, 1.0, 1.5, 2.0, 2.5 and 3.0 nm hydration shells. Each computation was performed for 101 frames of an MD simulation trajectory. Color scheme: red (lysozyme), green (myoglobin) and blue (RNase). Point representation: sphere (spherical average) and cube (FFT-based method).

common calculations. If the molecular system size exceeds 31.4 nm, the value of r in $N_{\text{FFT}} = r/r_{\text{bin}}$ becomes the dimension of the envelope. In this case, the computational time of our algorithm will increase for a larger system size (or a larger number of atoms). In addition, although we constructed protein-shaped envelopes to ensure a fair comparison with the spherical average method, the computational efficiency of our FFT-based method does not depend on the envelope shape for the tested cases. Simply using a spherical envelope will further reduce the computational cost spent to construct the envelope, which could be significant for large protein molecules.

We also implemented the FFT-based approach with the original *AXES* method, and tested the improvement of computational efficiency. Lysozyme was chosen as the benchmark system, and the calculated scattering intensity used 101 frames of a water box obtained from a constant NVT simulation. The computational efficiency was compared using our homemade implementation of both the FFT-based method and the Fibonacci grid method. The number of directions in the Fibonacci grid method was chosen to be 1589. The maximum distance between the protein surface and the water surface was chosen to be 0.3 nm and $\delta\rho$ was chosen to be 0.04 in our computation. It was found that the FFT-based method reduced the computational time from 1290 to 337 s for a system containing about 6500 atoms (including protein molecule, displaced water molecules and hydration water molecules).

3.3. Scattering profiles compared with experimental data

The computed scattering intensities of the three benchmark systems (lysozyme, myoglobin and RNase) were compared with the experimental data from SASBDB (<http://www.sasbdb.org>). We have neglected some less accurate experimental data points at low q and used only the scattering intensities starting from 0.309, 0.248 and 0.631 nm⁻¹, respectively, for the three proteins. This is because the Guinier plots of the three proteins' intensity data are nonlinear in the discarded regions. In the computations, the thickness of the water layer was 0.7 nm, consistent with the value presented in the literature (Park *et al.*, 2009; Chen & Hub, 2014). We also calculated the scattering profiles with the same water layer thickness, using the WAXSiS server (<http://waxsis.uni-goettingen.de>) (Chen & Hub, 2014). All the computed scattering profiles were first interpolated to have the same resolution as the experimental profiles and then fitted to the experimental data by minimizing χ^2 , defined in equation (8). The comparisons are shown in Fig. 3 and the χ^2 values are listed in Table 1. In Fig. 3, the computed profiles from the two methods fit well to the experimental curves. The χ^2 values for two out of three proteins are smaller than 1.0, indicating perfect agreement between the computation results and scattering experiments. For lysozyme, the computed scattering profiles from both methods agree very well with the experimental profile. The result of the WAXSiS server is slightly better than that of the FFT method, with a χ^2 value of 0.515 versus 0.564. It is observed that the main discrepancy between

Table 1

The χ^2 values for the scattering intensities computed by the WAXSiS server and our FFT-based method, using experimental scattering intensities as standard for the three benchmark proteins.

	Lysozyme	Myoglobin	RNase
WAXSiS	0.515	1.192	0.848
FFT	0.564	2.130	0.504

the experimental curves and the FFT method is in the range of $q \simeq 6\text{--}7\text{ nm}^{-1}$. The computed scattering profiles for myoglobin deviate most among the three proteins. The χ^2 value of the WAXSiS server is 1.192, and the computed curve fits to the experimental one reasonably until $q \simeq 8.5\text{ nm}^{-1}$. The result of the FFT method deviates slightly from the experimental one in the region of $q \simeq 2\text{--}3\text{ nm}^{-1}$ and $q \simeq 5.5\text{--}7.5\text{ nm}^{-1}$. The χ^2 value of 2.130 is also slightly larger than that of the WAXSiS server, while the overall shape of the computed profile is still in reasonable agreement with the experimental curve. One possible reason for the relatively large χ^2 value in this calculation is that the protein contains a large cofactor HEME. The porphyrin macrocycle results in a large number of delocalized π electrons, and the classical force field may not be able to describe the molecular interactions accurately. A significantly larger χ^2 value is obtained when the trajectory with backbone restraints is used to calculate the scattering intensity (supporting information, Table S1). This means that the initial crystal structure is quite far away from the protein conformation in solution. For RNase, the scattering profile computed

by the FFT method is slightly better than that obtained by the WAXSiS server, with χ^2 values of 0.504 compared with 0.848. Moreover, the computed curves from both methods fit the experimental profile very well. Starting from $q \simeq 8\text{ nm}^{-1}$, the scattering intensities computed by the WAXSiS server become a little higher than the experimental curve, whereas for the FFT-based method, there is a slight overestimation around $q \simeq 7\text{ nm}^{-1}$.

In conclusion, the results from the FFT-based method are comparable to those from the WAXSiS server, using experimental data as the standard. The small discrepancies between the results from the two methods may originate from the different conformations sampled during the explicit solvent MD simulations. We also compared our FFT results with those from our implementation of the spherical average method (data not shown), and only very small deviations in the high- q region were observed. Additionally, from Table S1, it is seen that the fluctuations of protein conformations contribute to the accuracy of scattering intensity calculations. This is consistent with the fact that the scattering intensity is the average result of multiple protein conformations in solution state. The contribution varies among proteins, depending on the difference between the initial structure and the solution conformation in the experimental environment.

Fig. 4 shows the comparison between the results for lysozyme from the AXES server and our FFT implementation. The χ^2 values between the computed curves and experimental data are close for the FFT-based method (0.611) and AXES (0.745). It should be noted that $\delta\rho = 0.04$ is used in our calculation instead of 0.02 used by the AXES server. We choose $\delta\rho$ to be 0.04 because the best fitting to experimental data is observed with this value. The difference of the water box trajectories used in the two calculations may account for the slight difference of the optimal $\delta\rho$ and the χ^2 values obtained.

We also find that the χ^2 values strongly depend on the fitting formula. As shown in Table S2 and Fig. S1, removing the

Figure 3

The scattering profiles computed using the all-atom simulation trajectories by the WAXSiS server and our FFT-based method, compared with the experimental scattering profiles after minimizing the χ^2 values.

Figure 4

The scattering profiles of lysozyme computed using the FFT implementation of the AXES approach and the AXES server, compared with the experimental scattering profile after minimizing the χ^2 values.

fitting parameter c substantially affects the fitting results of WAXSiS, making all the results even worse than those from the FFT method. Nevertheless, the FFT-based method can produce reasonable fitting results for all the cases tested.

3.4. Envelope size effect

To implement equation (2), the envelope size should be large enough to include a certain number of 'bulk-like' solvent molecules (Park *et al.*, 2009). In the previous studies (Chen & Hub, 2014; Park *et al.*, 2009) several small values of the shell thickness were tested, and the effects on scattering properties such as the solute radius of gyration and zero angle scattering intensity were investigated. Here, we extend the analysis to a wider range of water layer thicknesses, using the more efficient FFT-based algorithm. In particular, we examine the effect of envelope size on the quality of fitting to experimental data. Three benchmark proteins were selected, and shell thicknesses ranging from 0.5 to 3 nm were used to calculate the

Table 2

The χ^2 values between the computed and experimental scattering intensities for the three benchmark proteins using different solvation shell thicknesses.

Thickness of solvation shell (nm)	Lysozyme	Myoglobin	RNase
0.5	0.623	2.118	0.486
0.7	0.564	2.130	0.504
1.0	0.551	2.084	0.514
1.5	0.654	1.888	0.429
2.0	0.727	2.643	0.536
2.5	0.714	2.796	0.641
3.0	2.035	2.926	0.967

scattering intensities from the simulation trajectories. The computed scattering profiles for lysozyme are shown in Fig. 5. From Fig. 5(a), it is seen that the profiles from various thicknesses are quite similar at low q (smaller than 5 nm^{-1}), except the last curve. They become slightly different at high q , and no obvious convergence is observed by increasing the envelope size. It is observed that the scattering intensity calculated using the hydration shell thickness of 3.0 nm results in a significant difference of both $I(0)$ and the rest of the scattering intensity. This could be caused by errors in the bulk water density calculation, as the very large spherical envelope becomes close to the simulation box size, affecting the density correction on the bulk defined as the water molecules out of the envelope. The calculation results are affected by the sampling of water density fluctuations in the MD simulations, and these effects are reflected in the $I(0)$ values and the whole curves when we choose different envelope sizes (Chen & Hub, 2014; Park *et al.*, 2009). Nevertheless, we show in Fig. 5(b) that the chosen envelope size (except the largest one) has no significant influence on the fitting quality with the experimental data, given the error bars on the experimental intensities. The χ^2 values of the three benchmark proteins are shown in Table 2, which further illustrate that all the tested envelope sizes ranging from 0.5 to 2.5 nm can give reasonable fitting results. Although the results calculated using 1.0 or 1.5 nm fit slightly better with the experimental curves, these small differences can be neglected owing to existing errors in experimental data. Therefore, our tests confirm that the previously recommended water layer thickness of around 0.7 nm (Chen & Hub, 2014; Park *et al.*, 2009) is appropriate.

4. Conclusion

Despite the abundant SAXS/WAXS data in structural biology, the information contained in an experimental profile is usually insufficient for solving the detailed molecular structure. We present an FFT-based method that can accurately compute theoretical scattering profiles from MD simulation trajectories, helping to extract the useful structural information from experimental SAXS/near-WAXS data. Similar to the previous methods based on MD trajectories (Chen & Hub, 2014; Kofinger & Hummer, 2013; Park *et al.*, 2009), our approach requires no fitting parameters when computing the scattering profiles and merely relies on the physics-based molecular interactions. Moreover, we also implemented the

Figure 5
(a) The scattering profiles of lysozyme computed using different thicknesses of the solvation shell. (b) The comparison with the experimental scattering profile.

FFT approach with a computational scheme that requires no expensive MD simulations (Grishaev *et al.*, 2010).

While the FFT-based approach is already computationally efficient for protein solution systems with moderate sizes, the scaling feature of the algorithm may benefit potential usage on proteins with high molecular weights. Other possible future applications include studies on multiple protein systems and aggregation effects, as well as MD-based and SAXS-guided structure optimizations.

The source code and executable programs for our implementations of the spherical average method and the FFT-based method can be accessed at <http://www.ntu.edu.sg/home/lylu>. Additional sample input and output files are also available at the same address.

We also calculated the scattering intensity using MD simulation trajectories with positional restraints added to protein backbone atoms. The χ^2 values between the computed and experimental scattering profiles of the three benchmark proteins are listed in Table S1 of the supplementary material. The χ^2 values evaluated without parameter c are also computed and listed in Table S2. The comparison of computed and experimental scattering intensity under this fitting scheme is plotted in Fig. S1.

Acknowledgements

This research is supported by the Tier 1 grant RG45/14 and the Tier 3 grant MOE2012-T3-1-008 from the Ministry of Education, Singapore. The authors thank Hui Ji at the National University of Singapore for valuable discussions on FFT. Technical support from Jochen S. Hub at the Georg-August-University Göttingen on the WAXSiS web server is acknowledged. Moreover, we acknowledge Wuan Geok Saw and Gerhard Gruber at the Nanyang Technological University for assistance in analyzing scattering data. Finally, the authors appreciate the proofreading of our manuscript by Nalaparaju Anjaiah at the Nanyang Technological University.

References

- Bardhan, J., Park, S. & Makowski, L. (2009). *J. Appl. Cryst.* **42**, 932–943.
- Bell, J. A. (1999). *Protein Sci.* **8**, 2033–2040.
- Blanchet, C. E. & Svergun, D. I. (2013). *Annu. Rev. Phys. Chem.* **64**, 37–54.
- Castiglioni, P. (2014). *Wiley StatsRef: Statistics Reference Online*. John Wiley and Sons.
- Chen, P. C. & Hub, J. S. (2014). *Biophys. J.* **107**, 435–447.
- Chen, P. C. & Hub, J. S. (2015). *Biophys. J.* **108**, 2573–2584.
- Cromer, D. T. & Waber, J. T. (1965). *Acta Cryst.* **18**, 104–109.
- Fischetti, R. F., Rodi, D. J., Gore, D. B. & Makowski, L. (2004). *Chem. Biol.* **11**, 1431–1443.
- Graewert, M. A. & Svergun, D. I. (2013). *Curr. Opin. Struct. Biol.* **23**, 748–754.
- Grishaev, A., Guo, L. A., Irving, T. & Bax, A. (2010). *J. Am. Chem. Soc.* **132**, 15484–15486.
- Hekkelman, M. L., te Beek, T. A. H., Pettifer, S. R., Thorne, D., Attwood, T. K. & Vriend, G. (2010). *Nucleic Acids Res.* **38**, W719–W723.
- Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). *J. Chem. Phys.* **79**, 926–935.
- Koch, M. H. J., Vachette, P. & Svergun, D. I. (2003). *Q. Rev. Biophys.* **36**, 147–227.
- Kofinger, J. & Hummer, G. (2013). *Phys. Rev. E*, **87**, 052712.
- Konarev, P. V., Petoukhov, M. V., Volkov, V. V. & Svergun, D. I. (2006). *J. Appl. Cryst.* **39**, 277–286.
- Lipfert, J. & Doniach, S. (2007). *Annu. Rev. Biophys. Biomol. Struct.* **36**, 307–327.
- Liu, H., Hexemer, A. & Zwart, P. H. (2012). *J. Appl. Cryst.* **45**, 587–593.
- Makowski, L., Bardhan, J., Gore, D., Lal, J., Mandava, S., Park, S., Rodi, D. J., Ho, N. T., Ho, C. & Fischetti, R. F. (2011). *J. Mol. Biol.* **408**, 909–921.
- Makowski, L., Gore, D., Mandava, S., Minh, D., Park, S., Rodi, D. J. & Fischetti, R. F. (2011). *Biopolymers*, **95**, 531–542.
- Maurus, R., Bogumil, R., Nguyen, N. T., Mauk, A. G. & Brayer, G. (1998). *Biochem. J.* **332**, 67–74.
- Mertens, H. D. T. & Svergun, D. I. (2010). *J. Struct. Biol.* **172**, 128–141.
- Merzel, F. & Smith, J. C. (2002). *Acta Cryst.* **D58**, 242–249.
- Nguyen, H. T., Pabit, S. A., Meisburger, S. P., Pollack, L. & Case, D. A. (2014). *J. Chem. Phys.* **141**, 22D508.
- Niebling, S., Björling, A. & Westenhoff, S. (2014). *J. Appl. Cryst.* **47**, 1190–1198.
- Olds, D. P. & Duxbury, P. M. (2014). *J. Appl. Cryst.* **47**, 1077–1086.
- Oroguchi, T., Hashimoto, H., Shimizu, T., Sato, M. & Ikeguchi, M. (2009). *Biophys. J.* **96**, 2808–2822.
- Park, S., Bardhan, J. P., Roux, B. & Makowski, L. (2009). *J. Chem. Phys.* **130**, 134114.
- Poitevin, F., Orland, H., Doniach, S., Koehl, P. & Delarue, M. (2011). *Nucleic Acids Res.* **39**, W184–W189.
- Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., Hess, B. & Lindahl, E. (2013). *Bioinformatics*, **29**, 845–854.
- Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. A. (2007). *Q. Rev. Biophys.* **40**, 191–285.
- Ravikumar, K. M., Huang, W. & Yang, S. C. (2013). *Biophys. J.* **104**, 183a.
- Rey, F. A. & Dumas, C. (1984). *Biochimie*, **66**, 121–125.
- Schmidt-Rohr, K. (2007). *J. Appl. Cryst.* **40**, 16–25.
- Schneidman-Duhovny, D., Hammel, M. & Sali, A. (2010). *Nucleic Acids Res.* **38**, W540–W544.
- Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. (2013). *Biophys. J.* **105**, 962–974.
- Sorenson, J. M., Hura, G., Glaeser, R. M. & Head-Gordon, T. (2000). *J. Chem. Phys.* **113**, 9149–9161.
- Stovgaard, K., Andreetta, C., Ferkinghoff-Borg, J. & Hamelryck, T. (2010). *BMC Bioinformatics*, **11**, 429.
- Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.
- Svergun, D. I. & Koch, M. H. J. (2003). *Rep. Prog. Phys.* **66**, 1735–1782.
- Tainer, J. A. (2011). *Biophys. J.* **100**, 38a.
- Valentini, E., Kikhney, A. G., Previtali, G., Jeffries, C. M. & Svergun, D. I. (2015). *Nucleic Acids Res.* **43**, D357–D363.
- Yang, S. C., Park, S., Makowski, L. & Roux, B. (2009). *Biophys. J.* **96**, 4449–4463.
- Zheng, W. J. & Tekpinar, M. (2011). *Biophys. J.* **101**, 2981–2991.