Accurate optimization of amino acid form factors for computing small-angle X-ray scattering intensity of atomistic protein structures

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

Received 12 February 2016

Accepted 15 May 2016

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

Keywords: small-angle X-ray scattering; form factors; proteins; solution X-ray scattering; coarse graining.

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

^aSchool of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, 637551, Singapore, and ^bCenter for Proteomics and Department of Nutrition, Case Western Reserve University, 10900 Euclid Avenue, BRB 929, Cleveland, OH 44106-4988, USA. *Correspondence e-mail: llylu@ntu.edu.sg

Structure modelling *via* small-angle X-ray scattering (SAXS) data generally requires intensive computations of scattering intensity from any given biomolecular structure, where the accurate evaluation of SAXS profiles using coarse-grained (CG) methods is vital to improve computational efficiency. To date, most CG SAXS computing methods have been based on a single-bead-per-residue approximation but have neglected structural correlations between amino acids. To improve the accuracy of scattering calculations, accurate CG form factors of amino acids are now derived using a rigorous optimization strategy, termed electron-density matching (EDM), to best fit electron-density distributions of protein structures. This EDM method is compared with and tested against other CG SAXS computing methods, and the resulting CG SAXS profiles from EDM agree better with all-atom theoretical SAXS data. By including the protein hydration shell represented by explicit CG water molecules and the correction of protein excluded volume, the developed CG form factors also reproduce the selected experimental SAXS profiles with very small deviations. Taken together, these EDM-derived CG form factors present an accurate and efficient computational approach for SAXS computing, especially when higher molecular details (represented by the q range of the SAXS data) become necessary for effective structure modelling.

1. Introduction

Small-angle X-ray scattering (SAXS) is widely used to conduct protein structural characterizations in a solution state (Putnam *et al.*, 2007; Mertens & Svergun, 2010; Graewert & Svergun, 2013; Koch *et al.*, 2003; Svergun & Koch, 2003; Petoukhov & Svergun, 2007; Lipfert & Doniach, 2007; Chu & Hsiao, 2001; Doniach, 2001; Toft *et al.*, 2008; Hura *et al.*, 2009). Structural information encoded in a one-dimensional protein SAXS profile is an ensemble average of different protein conformations and orientations. Structural modelling from SAXS profiles generally requires sophisticated algorithms and iterative evaluations of scattering intensity from given model structures (Chacón *et al.*, 1998; Svergun, 1999; Svergun *et al.*, 2001; Förster *et al.*, 2008; Franke & Svergun, 2009; Schneidman-Duhovny *et al.*, 2012; Yang, 2014). In fact, there are already a few computing methods that accurately compute SAXS (or even near-wide-angle scattering) from an atomistically detailed structure model (Svergun *et al.*, 1995; Schneidman-Duhovny *et al.*, 2010, 2013; Grishaev *et al.*, 2010; Oroguchi *et al.*, 2009; Park *et al.*, 2009; Köfinger & Hummer, 2013; Chen & Hub, 2014; Nguyen *et al.*, 2014; Poitevin *et al.*, 2011; Olds & Duxbury, 2014; Liu *et al.*, 2012). However, the computations for large systems can be demanding, owing to the large number of evaluations required in a structure

© 2016 International Union of Crystallography

refinement. A simplified coarse-grained (CG) representation of protein structures can greatly improve the efficiency of SAXS profile evaluations, especially for structure-modelling applications.

Recently, a number of CG-based computing methods for evaluating SAXS profiles have been reported (Yang *et al.*, 2009; Stovgaard *et al.*, 2010; Zheng & Tekpinar, 2011; Ravikumar *et al.*, 2013; Schneidman-Duhovny *et al.*, 2013; Niebling *et al.*, 2014). These methods have been successful in terms of their mapping schemes (*i.e.* from an all-atom to CG representation), their parameterization of the form factors of simplified residues and their treatment of the hydration layer in the scattering. For example, *Fast-SAXS-Pro* (Yang *et al.*, 2009; Ravikumar *et al.*, 2013) uses a residue-based CG mapping scheme, where a single bead is placed at the C α position of each amino acid residue. Simplified form factors of residue beads are derived *via* a Debye summation over all the atoms contained within each residue, based on averaging over all possible conformations observed in a set of non-redundant proteins from the Protein Data Bank (PDB). We refer to this as the isolated bead approximation (IBA) method herein. Such a method is also termed the single-bead approximation (SBA) method by Niebling *et al.* (2014). Note that the hydration layer is represented by explicit CG water molecules near the protein surface in the IBA method, and it is shown that the resulting SAXS profiles are comparable to experimental data as well as those evaluated using *CRY SOL* (Svergun *et al.*, 1995). Similarly, Zheng & Tekpinar (2011) adopted a one-bead-per-residue CG scheme, except an implicit protein hydration shell was used by incorporating water molecules from the hydration layer into close protein residues. Very recently, Niebling *et al.* (2014) extended this IBA strategy to a finer CG scheme of one bead representing four heavy atoms, a common approach used in developing the MARTINI force field for molecular dynamics simulation (Marrink *et al.*, 2007). Note that no protein hydration layer scattering was considered in that approach. Nonetheless, it successfully reproduced the experimental scattering difference due to the conformational change of a protein, for a q range up to $0.75 \text{ \AA}^{-1}$. The success of that approach suggests that the hydration layer effects from the two protein conformations are similar.

In the most popular IBA or IBA-like approaches, the key step is based on the Debye formula, which computes the CG bead intensity individually, without considering the structural correlations between the beads in a protein. This leaves significant room for further improvement in developing more accurate CG form factors. In principle, such a structural correlation can be incorporated in a self-consistent fashion, though it is less popular than IBA because of its expensive iterative procedure (Niebling *et al.*, 2014). Alternatively, Stovgaard *et al.* (2010) used a maximum-likelihood-based Markov chain Monte Carlo method for the form-factor parameterization. In this work, the fitting targets were the theoretical SAXS intensities calculated by *CRY SOL* excluding the contribution of the protein hydration layer. CG SAXS profiles from that approach fit those calculated using

CRY SOL very well up to $q = 0.75 \text{ \AA}^{-1}$ for a two-bead-per-residue centre of mass mapping. The method of Stovgaard and co-workers is based on a computationally challenging nonlinear optimization, and a direct comparison with experimental data is absent because the hydration layer was not included.

In this paper, we introduce an optimization scheme, termed electron-density matching (EDM), to systematically parameterize the form factors of coarse-grained amino acids from their atomistic representations. This optimization is achieved *via* least-squares fitting of the Fourier transforms of their electron densities. More specifically, we generated a large set of non-redundant proteins from the PDB, and the Fourier transforms of the protein electron densities calculated from atomic form factors were used as fitting targets. Compared with a previous nonlinear optimization (Stovgaard *et al.*, 2010), this EDM optimization using a linear scheme provides a simple but mathematically guaranteed optimal solution. Moreover, the linearity of a Fourier transform allows us to separate its components so that the protein excluded-volume effect can be optimized separately. Specifically, the EDM method was tested for both a one-bead- and a two-bead-per-residue CG scheme. The resulting form factors are shown to reproduce all-atom SAXS profiles better than the IBA approach. Here, by including an explicit solvent hydration layer on the protein surface, we show that the EDM-derived CG form factors are able to compute theoretical SAXS profiles that accurately fit the corresponding experimental data.

2. Theory and computational details

2.1. Protein coarse-grained representations

Two types of CG mapping are implemented in this work, using, respectively, one and two beads to represent an amino acid residue. The position of each bead is placed at each residue's (or atom group's for the two-bead-per-residue case) centre of mass (COM) or centre of electrons (COE) (Zheng & Tekpinar, 2011), or alternatively the C α atom. Here, the COE scheme is used and represented *via* the following formula:

$$\mathbf{R}_{\text{CG}} = \frac{\sum_j e_j \mathbf{r}_j}{\sum_j e_j}, \quad (1)$$

where $\mathbf{R}$ and $\mathbf{r}_j$ are the CG and atomistic coordinates, respectively, e_j is the number of electrons in atom j , and j loops over all the atoms inside the CG bead. This CG scheme is proposed on the basis of the physical meaning of CG form factors, *i.e.* the Fourier transform of the electron density of a given CG bead. Using a form factor as a function of a scalar q is equivalent to assuming a spherically symmetric electron density in real space. Thus, placing the CG bead at the centre of electrons can produce a good overlap between the atomistic and CG electron-density distributions. When the COE CG scheme is implemented for the one-bead model, each bead is placed at the COE of its corresponding residue. For the two-bead model, two CG sites are used: one for the backbone

containing six atoms (N, C α , C, O, HN and H α) and the other for its side chain which consists of the rest of the atoms in the residue. For the COE mapping scheme, these two beads are placed at their corresponding COEs, resulting in a total of 21 different types of CG bead, *i.e.* one backbone bead and 20 side-chain beads. Besides the COE type of mapping, given the models' conceptual simplicity, a simple C α -type one-bead model is also implemented, as well as a two-bead model of a C α backbone plus a COM side chain. The results of different types of CG schemes are compared in our computations.

2.2. Derivation of coarse-grained form factors

2.2.1. Electron-density matching. Here, an optimization scheme, termed the EDM method, is used to derive CG form factors of amino acid residues from their atomistic structure coordinates. Specifically, this EDM optimization is achieved by minimizing the difference between the two three-dimensional Fourier transforms of protein electron density which are calculated from the CG and atomistic coordinates of the same protein. Compared with the previous fitting based on the one-dimensional scattering intensity $I(q)$ (Stovgaard *et al.*, 2010), this electron-density-based matching is a more rigorous criterion since it contains information prior to the orientational averaging in computing $I(q)$. In principle, such an EDM method is capable of faithfully reproducing the electron-density distribution from the atomistic structure, if the Fourier transform of the protein electron density is fitted properly over all scattering vectors $\mathbf{q}$.

Specifically, from a given atomistically detailed structure, the Fourier transformation of protein electron density at vector $\mathbf{q}$ can be calculated by

$$A(\mathbf{q}) = \sum_j f_j(q) \exp(-i\mathbf{q} \cdot \mathbf{r}_j), \quad (2)$$

where j loops over all the atoms within the protein, $\mathbf{r}_j$ are the atomic coordinates of the j th atom and $f_j(q)$ is the atomic form factor of the j th atom. The atomic form factor $f_j(q)$ can be evaluated using the following four-Gaussian approximation:

$$f_j(q) = \sum_{k=1}^4 a_k \exp[-b_k(q/4\pi)^2] + c, \quad (3)$$

where a_k , b_k and c are the Cromer–Mann parameters (Cromer & Waber, 1965) obtained from *International Tables for Crystallography* (Brown *et al.*, 2006). To incorporate the effect of displaced solvent, we adopt the modified atomic form factor (Fraser *et al.*, 1978) according to the following equation:

$$f_j^{\text{excl}}(q) = f_j(q) - v_j \rho_0 \exp(-\pi v_j^{2/3} q^2), \quad (4)$$

where v_j is the van der Waals volume of atom j and ρ_0 is the electron density of the bulk water (334 e nm $^{-3}$). Thus, in SAXS computations using equation (2), the atomic form factors $f_j(q)$ are actually replaced by $f_j^{\text{excl}}(q)$.

On the other hand, the same Fourier transform $A(\mathbf{q})$ in equation (2) can be computed but using CG form factors,

$$A(\mathbf{q}) = \sum_j F_j^{\text{CG}}(q) \exp(-i\mathbf{q} \cdot \mathbf{R}_j^{\text{CG}}), \quad (5)$$

where J loops over all CG beads and $F_j^{\text{CG}}(q)$ is the CG form factor of the J th bead.

It is not technically possible to fit $A(\mathbf{q})$ for the entire continuous reciprocal $\mathbf{q}$ space, so a discrete set of $M = 1000$ directions was selected for each scalar value of q , in a similar spirit to that used in a spherical average method for computing scattering intensities (Park *et al.*, 2009; Chen & Hub, 2014). The directions are chosen according to a known spiral method (Ponti, 1999), where the spherical coordinates of each direction, the θ and φ angles, are determined by

$$\begin{aligned} \theta_m &= \arccos \frac{2m-1-M}{M}, \\ \varphi_m &= (\pi M)^{1/2} \arcsin \frac{2m-1-M}{M}. \end{aligned} \quad (6)$$

Ideally, the Fourier transform $A(\mathbf{q})$ of an arbitrary protein calculated from CG form factors should be equal to that from atomic form factors. Fortunately, this requirement can be achieved mathematically by combining equation (2) and equation (5). Specifically, this optimization between CG and atomistic form factors is achieved *via* the following linear overdetermined equations, where the number of equations available is more than the number of unknown variables to be solved, for a set of vectors $\mathbf{q}$ and a selected protein structure:

$$\begin{aligned} &\begin{bmatrix} \sum_j^{N_1} \exp(-i\mathbf{q}_1 \cdot \mathbf{R}_j) & \cdots & \sum_j^{N_t} \exp(-i\mathbf{q}_1 \cdot \mathbf{R}_j) \\ \vdots & \ddots & \vdots \\ \sum_j^{N_1} \exp(-i\mathbf{q}_M \cdot \mathbf{R}_j) & \cdots & \sum_j^{N_t} \exp(-i\mathbf{q}_M \cdot \mathbf{R}_j) \end{bmatrix} \begin{bmatrix} F_1^{\text{CG}}(q) \\ \vdots \\ F_t^{\text{CG}}(q) \end{bmatrix} \\ &= \begin{bmatrix} \sum_j^{N_a} f_j^{\text{excl}}(q) \exp(-i\mathbf{q}_1 \cdot \mathbf{r}_j) \\ \vdots \\ \sum_j^{N_a} f_j^{\text{excl}}(q) \exp(-i\mathbf{q}_M \cdot \mathbf{r}_j) \end{bmatrix}. \end{aligned} \quad (7)$$

Here t is the number of CG site types that describe different types of residue, *i.e.* $t = 20$ for a one-bead-per-residue model and $t = 21$ for a two-bead-per-residue model. N_1 to N_t represent the number of CG beads for each CG site type, and N_a is the number of atoms in the selected protein structure. $\mathbf{R}_j$ are the CG coordinates and $\mathbf{r}_j$ are the atomic coordinates. The vectors $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_M$ represent a total of $M = 1000$ different directions for a pre-defined scalar q range, which is set to 0–0.5 Å $^{-1}$ in this work. In principle, the EDM algorithm allows optimization over a broad q range, though the upper limit of q (q_{max}) is often determined by the resolution of a chosen CG simplification scheme in order to avoid over-fitting, *e.g.* a very high q_{max} for a very coarse-grained representation.

To include multiple protein structures in the EDM optimization process, we simply apply equation (7) to each protein, so the entire set of protein structures from our database is described by

$$\begin{aligned} \mathbf{A}_1 \mathbf{f} &= \mathbf{b}_1 \\ \mathbf{A}_2 \mathbf{f} &= \mathbf{b}_2 \\ &\vdots \\ \mathbf{A}_P \mathbf{f} &= \mathbf{b}_P \end{aligned} \quad \text{or} \quad \mathbf{A} \mathbf{f} = \mathbf{b}. \quad (8)$$

Here P is the number of protein structures in the data set, and each row is applied to each protein, equivalent to equation (7). Matrices $\mathbf{A}$ and $\mathbf{b}$ are constructed by $\mathbf{A} = (\mathbf{A}_1^T \mathbf{A}_2^T \dots \mathbf{A}_P^T)^T$ and $\mathbf{b} = (\mathbf{b}_1^T \mathbf{b}_2^T \dots \mathbf{b}_P^T)^T$. It is technically unrealistic to construct a set of CG form factors to satisfy equation (8) completely, so equation (8) is solved in a least-squares manner. We can rewrite equation (8) in a normal equation form:

$$\mathbf{A}^T \mathbf{A} \mathbf{f} = \mathbf{A}^T \mathbf{b} \quad \text{or} \quad \mathbf{G} \mathbf{f} = \mathbf{d}, \quad (9)$$

where the matrix $\mathbf{G}$ is the normal matrix of the original matrix $\mathbf{A}$ in equation (8), and the vector $\mathbf{d}$ is $\mathbf{A}^T \mathbf{b}$. Equations (8) and (9) can be solved by many numerical methods, and in this work we applied singular-value decomposition (SVD) (Golub & Kahan, 1965) to solve the matrix equations. Thus, best-fit CG form factors are derived from the entire set of protein structures *via* Fourier transforms of atomistic electron-density distributions.

2.2.2. Isolated-bead approximation. To make a comparison with the EDM method, the IBA, which has been widely used in earlier work (Yang *et al.*, 2009; Zheng & Tekpinar, 2011; Ravikumar *et al.*, 2013; Niebling *et al.*, 2014), was also implemented here. In the IBA, a CG form factor is determined by calculating the square root of the scattering intensity of all the atoms within each CG bead *via* averaging over all possible CG group conformations in a selected protein data set by

$$F^{\text{CG}}(q) = \left\langle \left[\sum_{i,j}^{N_{\text{CG}}} f_i^{\text{excl}}(q) f_j^{\text{excl}}(q) \frac{\sin(qr_{ij})}{qr_{ij}} \right]^{1/2} \right\rangle_{\text{PDB}}, \quad (10)$$

where i and j loop over all the atoms within each CG bead and the ensemble average is performed over all possible conformations of the corresponding atom group in the same set of 1003 protein structures as used in the EDM optimization.

As pointed out previously by Niebling *et al.* (2014), equation (10) will only produce positive form factors. Specifically, the one-bead-per-residue form factors from the original IBA work are all positive, though a few negative form factors appear for less coarse grained models, owing to the inclusion of the excluded-volume terms. When corrected by equation (4), the atomic form factor of hydrogen atoms becomes negative (-0.72) at $q = 0$. The form factor of a CG bead at $q = 0$ should be equal to the summation of all the atomic form factors inside the CG bead:

$$F^{\text{CG}}(q = 0) = \sum_j^{N_{\text{CG}}} f_j(q = 0). \quad (11)$$

Thus, the CG form factors of those beads that contain many hydrogen atoms could be negative $F(q)$ values at $q = 0$. In our CG scheme, negative $F^{\text{CG}}(q = 0)$ values only occur for some side-chain beads of the two-bead model, and equation (11) is not satisfied for these types of bead. Thus, we performed

corrections on the affected CG form factors (Niebling *et al.*, 2014).

2.3. Modelling experimental SAXS profiles

2.3.1. Contribution of the protein hydration layer. To validate the EDM-derived CG form factors using experimental SAXS data, the protein hydration shell is represented by a layer of explicit water molecules, the coordinates of which are determined by the following method.

We generated a pre-equilibrated cubic water box from a 1 ns constant NPT (particle number–temperature–pressure) molecular dynamics simulation at 300 K and 1 bar (1 bar = 100 kPa). The water model used in the simulation was SPC/E (Berendsen *et al.*, 1987) and the size of the final box was 9.9852 nm. Additional periodic images of the water box were extended to increase the box size when necessary. Each protein molecule was overlapped with the pre-equilibrated cubic water box to produce the solvated protein system. If the minimum distance between a water oxygen atom and a protein atom was larger than $r_{\text{vdw}} + r_{\text{wat}}$ and smaller than $r_{\text{vdw}} + r_{\text{wat}} + 3 \text{ \AA}$, the water molecule was accepted as part of the protein hydration layer. Here, r_{vdw} is the van der Waals radius of the nearest protein atom to the water oxygen atom and r_{wat} is the effective radius of a water molecule, defined to be 1.4 Å. If the minimum distance was larger than r_{vdw} and smaller than $r_{\text{vdw}} + r_{\text{wat}}$, the water molecule was considered to be partially overlapped with protein atoms. Thus, we used a stochastic procedure to determine whether this water molecule was accepted or not, following the previous work by Grishaev *et al.* (2010). The probability of including this type of water molecule in the protein hydration layer was calculated by

$$P_{\text{accept}} = \prod_j \left(\frac{r_j - r_{\text{vdw}}}{r_{\text{wat}}} \right)^\gamma, \quad (12)$$

where j loops over all protein atoms whose distances r_j to the water O atom are closer than $r_{\text{vdw}} + r_{\text{wat}}$. The parameter γ was empirically chosen to be 0.1 following Grishaev *et al.* (2010). Here, the basic idea is that if a certain water molecule overlaps with more protein atoms, the tendency to be regarded as displaced water instead of hydration water is higher. Finally, a post-processing procedure was performed in order to remove water molecules in the small cavities within the protein molecule. We removed water molecules that had less than two neighbours within 3.8 Å in the hydration water set. This procedure was done iteratively until no more water molecules were removed.

The hydration water molecules are included in the calculation of the CG SAXS profile by placing CG beads at the O atoms. The CG form factor of CG water molecules is obtained using IBA, as follows:

$$F_w^{\text{CG}}(q) = \left[\sum_{i,j}^3 f_i(q) f_j(q) \frac{\sin(qr_{ij})}{qr_{ij}} \right]^{1/2}, \quad (13)$$

where the r_{ij} values are taken from the standard SPC/E water model. The atomic form factors for the water O and H atoms

are modified to consider the electron-withdrawing effect following Sorenson *et al.* (2000). When computing the CG SAXS profile, the CG form factor of water is further scaled by a weight factor w_{wat} to consider the excess electron density of hydration water compared with bulk water [*i.e.* $F_{\text{w}}^{\text{CG}}(q) = w_{\text{wat}} F_{\text{w}}^{\text{CG}}(q)$] (Ravikumar *et al.*, 2013).

2.3.2. Scalable excluded volume. It is reported in the literature that sometimes the Gaussian sphere approximation will not correctly reproduce the protein excluded volume (Fraser *et al.*, 1978). As a result, computational tools like *CRY SOL* can tune the excluded-volume term in order to obtain optimal fitting results with experimental SAXS profiles. Similarly, we also decompose the CG form factors parameterized by EDM into two terms, one for protein atoms only and another considering the excluded-volume effect. After the first round of EDM optimization, the over-determined equation (8) is constructed from the same protein data set again, using the original atomic form factors in equation (3) without considering the excluded-volume effect. The resulting CG form factors from the second round of parameterization only include the effect of protein atoms and are denoted as $F_{\text{vacuum}}^{\text{CG}}$. We subsequently obtain the term responsible for the excluded-volume effect by subtracting the previously obtained CG form factors from $F_{\text{vacuum}}^{\text{CG}}$, *i.e.* $F_{\text{excl}}^{\text{CG}} = F_{\text{vacuum}}^{\text{CG}} - F^{\text{CG}}$. After this, we can tune the weight of the excluded-volume term to obtain a modified set of CG form factors as $F^{\text{CG}'} = F_{\text{vacuum}}^{\text{CG}} - w_{\text{excl}} F_{\text{excl}}^{\text{CG}}$. As shown in §3.5, this process generates better fitting results to experimental profiles.

2.3.3. Optimization of the contributions of the hydration layer and excluded volume. As mentioned above, there are two free parameters w_{wat} and w_{excl} in the EDM modelling of experimental profiles, which can be tuned to fit experimental SAXS data. As the optimization target is the χ^2 value between calculated and experimental SAXS profiles, the analytical form of the derivatives is in general not available. So we apply the Nelder–Mead simplex algorithm to optimize the two parameters (Nelder & Mead, 1965). The initial values were set to be $w_{\text{wat}} = 0.03$ and $w_{\text{excl}} = 1.0$. The optimization converges when the size of the simplex is smaller than 10^{-3} .

2.4. Selection of protein data set

We constructed a high-resolution non-redundant protein data set using the *PISCES* server (Wang & Dunbrack, 2003). Crystal structures with a resolution higher than 1.8 Å, an *R* factor smaller than 0.15 and no missing residues were chosen. The PDB structures were culled by chains, and the sequence identity was filtered to be smaller than 10% to remove redundancy. This resulted in the 1003 single-domain structures in our data set (as of June 2015). The missing side-chain residues were completed using the function module *complete_pdb* in *Modeller* (Šali & Blundell, 1993). Structure clashes were further removed by performing a short energy minimization using *Gromacs 4.5.5* (Pronk *et al.*, 2013) and the *CHARMM* force field (MacKerell *et al.*, 1998).

2.5. Comparison of SAXS profiles

To measure the discrepancy between SAXS profiles, a scoring function χ^2 is defined as follows:

$$\chi^2 = N_q^{-1} \sum_{i=1}^{N_q} \left[\frac{I_{\text{exp}}(q_i) - f I_{\text{CG}}(q_i)}{\sigma(q_i)} \right]^2, \quad (14)$$

where N_q is the number of data points and $I_{\text{exp}}(q_i)$ are the experimental SAXS data. The parameter f is used to scale the CG SAXS profile to fit the target profile, and $\sigma(q_i)$ are the standard errors in $I_{\text{exp}}(q_i)$. Interpolation of $I_{\text{CG}}(q_i)$ by means of cubic splines is performed so that an identical set of q points is used in the two sets of scattering intensities for an effective χ^2 calculation.

In the case where a calculated all-atom SAXS profile is used as a reference to assess the performance of a CG method, an analogue of equation (14) is used to measure the difference between two SAXS profiles. The scaling parameter f is set to be 1.0 in this case, as both curves are from theoretical predictions. As a result, we define the scoring function

$$R = N_q^{-1} \sum_{i=1}^{N_q} \left[\frac{I_{\text{AA}}(q_i) - I_{\text{CG}}(q_i)}{\sigma(q_i)} \right]^2, \quad (15)$$

where $I_{\text{AA}}(q_i)$ is the atomistic SAXS profile used as the target. Additionally, because $\sigma(q_i)$ are not available for theoretical curves, an empirical standard error is adopted (Stovgaard *et al.*, 2010) as follows:

$$\sigma(q_i) = I(q_i)(q_i + \alpha)\beta, \quad (16)$$

where $\alpha = 0.15$ and $\beta = 0.3$. In general, the values of the R score are smaller than the typical χ^2 in equation (14), as theoretical SAXS profiles lack experimental fluctuations.

3. Results and discussion

3.1. Amino acid form factors

To calculate SAXS intensity efficiently for a given structure model, CG form factors of amino acids are developed by accurately accounting for the electron-density distributions of amino acids. The EDM method is used here to derive accurate form factors for amino acids. The results of our EDM-based scattering computations are further compared with those from the IBA method.

Note that this EDM method is essentially inspired by the force-matching approach that has been used in molecular modelling (Izvekov & Voth, 2005; Noid *et al.*, 2008); here, the EDM aims to reproduce the electron-density distributions of given protein structures. This goal is achieved by fitting Fourier transforms of electron density computed from atomistic structures [equation (7) and §2.2.1] to derive a set of simplified or CG form factors of amino acids for accurate SAXS computation.

Two specific CG mapping schemes are implemented: a one-bead-per-residue and a two-bead-per-residue representation. In the one-bead case, the location of the C α atom or the COE of the entire residue was used to position its corresponding

bead. Similarly, in the two-bead-per-residue scheme, backbone and side-chain beads were represented by the position of the C α atom and the centre of mass (COM) of the entire side chain, or the COE of each backbone/side-chain group, depending on the CG model used. For both types of CG scheme, CG form factors were evaluated using both the EDM method and the IBA method, which results in a total of six sets of CG form factors (see Table 1). The resulting form factors are shown in Fig. 1 (and supplementary Figs. S1 and S2).

Fig. 1 shows the CG form factors for amino acids obtained for a two-bead-per-residue scheme, where the COE is used for the positioning of both the backbone and side chain. The results of the EDM method and the IBA are shown in Figs. 1(b) and 1(c), respectively. It is not surprising that the form factors of the backbone and side-chain beads are quite different, owing to the larger variation in the side chains. Specifically, the EDM results show that the side-chain form factors generally have a sharp increase in the q range ~ 0 – $0.3 \text{ \AA}^{-1}$, though some of them reach a maximum value at $q \simeq 0.3 \text{ \AA}^{-1}$ and start to decrease after this point. However, the

Figure 1

The CG form factors obtained from EDM and IBA. (a) Illustration of a two-bead-per-residue CG scheme, where each bead is placed at its COE. For the illustrative structure shown in sticks, the backbone and side chain of each amino acid are coarse grained into by blue and red beads. (b), (c) The resulting CG form factors from EDM and IBA, respectively, where the blue and red curves represent form factors of the backbone and side-chain beads of all 20 residues, respectively. The ordering of their side-chain form factors is as follows, based on their intensity at $q = 0$: Arg, His, Asp, Asn, Glu, Cys, Gln, Met, Trp, Tyr, Ser, Thr, Lys, Gly, Phe, Ala, Pro, Val, Leu and Ile. Additional results from using other CG approaches can be found in Figs. S1 and S2 of the supporting information.

Table 1

A list of the six sets of coarse-grained form factors computed here and their corresponding figures.

For IBA, the computation of form factors is not related to the coarse-grained bead positions, so there are only two sets of IBA form factors.

	EDM	IBA
Two-bead COE	Fig. 1(b)	Fig. 1(c)
Two-bead C α + COM	Fig. S2(c)	Fig. 1(c)
One-bead COE	Fig. S1(b)	Fig. S1(c)
One-bead C α	Fig. S2(d)	Fig. S1(c)

IBA-derived form factors tend to have a continuous increase over the entire q range, with their maximum values appearing at $q > 0.5 \text{ \AA}^{-1}$ (data not shown). On the other hand, for the backbone form factors, the EDM results show a local minimum at $q \simeq 0.3 \text{ \AA}^{-1}$, while the IBA results yield a flat distribution around this q range.

The differences in form factors between the EDM and IBA methods may originate from the way that different beads are aligned and orientated with respect to each other and the way that their orientations are correlated. In IBA, form factors are computed separately for each type of CG bead, assuming that their rotations are completely independent and not biased to any particular orientation, so no bead-to-bead correlation is taken into account. However, in EDM, the entire set of CG beads is used to fit the Fourier transform of the electron density of the protein. As such, the correlations between different CG beads, represented by the off-diagonal elements of matrix **G** in equation (9), are considered explicitly.

Apart from the two-bead COE form factors in Fig. 1, the CG form factors using other CG schemes are presented in the supporting information (Figs. S1 and S2). In all the mapping cases, the EDM results are very different from the IBA counterparts. Our optimized form factors are also different from those in the previous work based on a nonlinear fitting of atomistic $I(q)$ values (Stovgaard *et al.*, 2010). This discrepancy may be due to the different optimization targets and numerical algorithms.

3.2. Accuracy and efficiency of amino acid form factors

To evaluate the accuracy of these derived CG form factors, SAXS profiles were computed for various proteins structures. These CG-based results were further compared with those computed from atomic form factors. A set of form factors is considered to be more accurate if the differences between the CG and all-atom scattering profiles are smaller. For comparison between different CG methods, the results from the EDM method are compared with the IBA approach, as well as another method by Stovgaard *et al.* (2010).

First, the accuracy of the EDM method is compared with that of the IBA method. The deviations between coarse-grained and all-atom scattering intensities are measured by R values computed using equation (15), and the distribution of these R values for the protein database is used to measure the accuracy of the EDM/IBA form factors. Figs. 2(a) and 2(b) list the R -value distributions of the 1003 high-resolution struc-

tures in our data set, for the two-bead and one-bead models, respectively. By comparing the two EDM distributions (from COE and from $C\alpha$ + COM mappings) (in red) with those from IBA (in blue), it is clear that the CG SAXS profiles match the all-atom profiles much better when using the EDM method. In Fig. 2(a), for EDM, the R values for 97.2 and 91.2% of proteins fall below 0.1 for the two two-bead CG schemes. As a comparison, the R values for only 6.08 and 5.58% of proteins fall below 0.1 for the IBA method. Additionally, 16.8 and 4.6% of protein SAXS profiles from IBA have R values greater than 0.4 for COE and $C\alpha$ + COM, respectively. Similar trends are also observed in the one-bead case in Fig. 2(b), though the deviations of both the EDM and the IBA method are in general greater than those from the two-bead mappings. The R distributions computed using $q \leq 0.35 \text{ \AA}^{-1}$ are listed in

Figure 2

The distribution of R values describing the difference in SAXS profiles between CG and all-atom models. SAXS profiles within a q range of ~ 0 – $0.5 \text{ \AA}^{-1}$ were used in the computation. Two different CG schemes are shown: (a) the two-bead-per-residue model and (b) the one-bead-per-residue model. The results of EDM and IBA are shown as red and blue curves, respectively, where circles indicate that the position of each CG bead is placed at its corresponding COE, and squares indicate that the position of each bead is placed at its $C\alpha$ atom (in the case of a one-bead-per-residue model) or at the $C\alpha$ atom for its backbone and the COM of its side chain (in the case of a two-bead-per-residue model).

supplementary Fig. S3, which illustrates the accuracy in the low- q region. Clearly, the EDM method performs better than IBA in this low- q region in both the two-bead and the one-bead case.

One significant aspect of the EDM method is the consideration of the orientational correlations between backbone and backbone, backbone and side chain, or side chain and side chain, which are now taken into account in EDM with improved accuracy. This orientational heterogeneity has been ignored in the IBA method, which uses the Debye formula and thus makes an implicit assumption of free rotations. This assumption is reasonable for very low q SAXS computing (e.g. $q < 0.2 \text{ \AA}^{-1}$), but not for higher q regions where more molecular details are required, as generally shown in high-resolution protein structures deposited in the PDB.

Apart from different parameterization methods, different CG mapping schemes will also lead to slightly different R -value distributions, as shown in Fig. 2. With the EDM method, the COE type of mapping always generates the optimal form factors, for both one-bead- and two-bead-per-residue cases. However, for IBA, the two-bead $C\alpha$ + COM scheme is slightly better than COE, while the one-bead $C\alpha$ scheme generates significantly more accurate results than COE. COE is supposed to be a very good choice for the EDM method, as it aims to fit atomistic electron densities. In terms of IBA, the choice of bead location becomes subtle as we cannot determine the optimal bead location. Although Fig. 2 suggests that $C\alpha$ -based mappings are preferred with IBA, supplementary Fig. S3(b) shows that COE is a better choice for the one-bead case if we only consider the region $q \leq 0.35 \text{ \AA}^{-1}$.

Another important aspect is the q range used in coarse-grained SAXS calculations, because most experimental data are in the range $q < 0.5 \text{ \AA}^{-1}$. To assess the effect of the q range, we computed the average R values of all the proteins in our database with different maximum q values q_{cut} . The CG form factors obtained from the EDM method with one- and two-bead COE CG schemes were applied, and the results are plotted in Fig. 3. Both curves show an increasing trend, when higher- q regions are used for R -value computation. Note that the error of the one-bead model is acceptable up to the q_{cut} value of around $0.35 \text{ \AA}^{-1}$, if an average R value of 0.1 is allowed. This reasonable performance suggests that the one-bead model is appropriate to compute scattering intensities for low- q regions, though for a larger q range a two-bead model becomes necessary for accuracy.

Solution X-ray scattering is also widely used to study the structure of intrinsically disordered proteins (IDPs). Thus, it is of great interest to test whether our EDM form factors parameterized from native protein structures work for these disordered proteins. We selected six NMR ensembles of IDPs as test cases, in total 100 structure models. The PDB codes of these IDPs are 1a5r (Bayer *et al.*, 1998), 1r8u (De Guzman *et al.*, 2004), 2fft (Song *et al.*, 2006), 2l14 (Lee *et al.*, 2010), 2lj1 (Singarapu *et al.*, 2011) and 2ly4 (Rowell *et al.*, 2012). The radii of gyration of these structure models range from 1.4 to 4.3 nm. Supplementary Fig. S4 illustrates the distribution of R values using the two-bead COE scheme. It was found that the SAXS

profiles computed from the EDM form factors can accurately reproduce the atomistic profiles, as all R values for the 100 NMR structures are smaller than 0.1 and the maximum value of the histogram is located at $R = 0.02$ – 0.03 . Supplementary Table S7 further shows that the EDM form factors with one-bead COE mapping can also describe the SAXS profiles of the selected IDPs with reasonable precision, though the errors are in general slightly higher than for the two-bead COE case. These results indicate that our EDM form factors can be applied beyond protein native structures.

In addition, the fitting results from the EDM method were compared with another parameterization method reported by Stovgaard *et al.* (2010). The CG form factors in their work were obtained by a maximum likelihood method fitting scattering intensities from *CRY SOL*. The protein data set from the original work was used, which contains 50 proteins. The S values defined by Stovgaard and co-workers, which are approximately the square roots of the R values, were evaluated following their approach. The distributions of the S values are compared in supplementary Fig. S5. Overall, the result of the EDM method is slightly better in terms of S values, though both methods yield accurate calculations of the SAXS intensities of all 50 proteins in the data set.

To evaluate the improvement in computational efficiency by using the CG form factors, a large protein GroEL:GroES2 complex (PDB code 4pkn; Fei *et al.*, 2014) that contains 8689 residues and 130 486 atoms was selected as a benchmark system. The computation of the SAXS profile using the two-bead CG scheme took 25 s, while the atomistic calculation for the same system required 203 s. This eightfold improvement was observed on an Intel Xeon CPU (E7-8850 at 2.00 GHz) with our in-house code for the spherical averaging algorithm (Park *et al.*, 2009). The computational efficiency can be further improved by using one-bead CG schemes with a slight trade-off of computational accuracy. Thus, the CG form factors

obtained using our EDM method can increase the computational efficiency while keeping a reasonable accuracy of the resulting SAXS profiles.

The above benchmark is only for the protein part of the scattering. The overall computational efficiency of SAXS modelling also depends on the calculation of hydration scattering, and various approaches are available in the literature. In this work, the hydration layer is modelled with explicit water molecules, following the work of Grishaev *et al.* (2010). However, our developed form factors can in principle be combined with other hydration scattering approaches.

The computational efficiency improvement given by CG form factors could be desirable for some situations where SAXS modelling is performed a large number of times (*e.g.* SAXS-based protein structure prediction). In the literature, CG molecular models have been implemented in a recently released SAXS-based structure prediction program *SCT* (Wright & Perkins, 2015). Compared with atomistic approaches, the CG-based method may require a backwards mapping from a CG to an all-atom structure. For protein systems, such a backwards mapping can be handled by existing software for a $\text{C}\alpha$ -based or arbitrary mapping (Li & Zhang, 2009; Wassenaar *et al.*, 2014). Additionally, CG form factors can be implemented for SAXS-guided molecular dynamics simulations, as CG molecular models are widely used to speed up molecular dynamics simulations (Niebling *et al.*, 2014). For the two-bead COE coarse-grained scheme that generates the best fitting results in this work, simple structure-based potentials may be used to provide conformational restraints during the structure modelling process, similar to the work by Zheng & Tekpinar (2011). Moreover, CG force fields parameterized from standard multiscale CG methods (Reith *et al.*, 2003; Lyubartsev & Laaksonen, 1995; Izvekov & Voth, 2005; Noid *et al.*, 2008) or statistical approaches (Miyazawa & Jernigan, 1985, 1996; Sippl, 1995; Jernigan & Bahar, 1996) can be developed for SAXS-guided CG molecular dynamics simulations with the set of CG form factors.

3.3. Example scattering profiles

To evaluate the quality of SAXS computing using the derived form factors, SAXS calculations using atomistic and CG form factors were performed for four different proteins of different sizes. The corresponding PDB codes are 4co8 (98 residues; Cooper *et al.*, 2015), 4jl5 (203 residues; Kerns *et al.*, 2015), 4uyb (400 residues; Kopec *et al.*, in preparation) and 1n62 (804 residues; Dobbek *et al.*, 2002). Figs. 4 and 5 show the results from the two-bead-per-residue and one-bead-per-residue CG schemes, respectively. As can be seen in Fig. 4, CG form factors from EDM generate better fitting results for all of the four selected proteins. For instance, for the two proteins in Figs. 4(a) and 4(b), the intensities from IBA with the COE mapping clearly deviate from the all-atom values in the intermediate q range (0.2 – $0.3 \text{ \AA}^{-1}$), while the EDM form factors fit the four all-atom curves almost perfectly. Interestingly, for the two-bead IBA results, the $\text{C}\alpha + \text{COM}$ (one for the backbone and the other for the side chain) mapping is

Figure 3

Average R values between CG and all-atom SAXS profiles as a function of q_{cut} values. The R values were evaluated over a q range below q_{cut} and averaged over all 1003 protein structures. The EDM method was used for the derivation of the CG form factors. The results from the one-bead and two-bead COE CG schemes are shown by blue and red curves, respectively.

slightly better than the COE counterpart for three of the four proteins, which is consistent with the previous discussion based on Fig. 2. A similar conclusion for the EDM/IBA comparison can be reached for the one-bead model results in Fig. 5, where the IBA curves deviate more obviously from the atomistic ones. We also note that, for the one-bead $C\alpha$ case with IBA, two of the four protein scattering profiles (Figs. 5*a* and 5*b*) show significant deviation in the intermediate q range $0.2\text{--}0.3\text{ \AA}^{-1}$, while the other two profiles (Figs. 5*c* and 5*d*) also show small deviations in the same q range. For comparison, the deviations for the one-bead COE case with IBA mainly appear in the higher q range. These results suggest that, for the IBA one-bead case, the COE bead location is better than $C\alpha$ for modelling experimental data with low $q \leq 0.35\text{ \AA}^{-1}$, while $C\alpha$ is a better choice for data with $q \leq 0.5\text{ \AA}^{-1}$. This conclusion is also supported by the R -value distributions in Fig. 2 ($q \leq 0.5\text{ \AA}^{-1}$) and supplementary Fig. S3 ($q \leq 0.35\text{ \AA}^{-1}$). Finally, the comparison between Figs. 4 and 5 demonstrates that the IBA method serves as a reasonable approximation for the less coarse grained two-bead-per-residue models, while the

improvement from the EDM approach becomes more significant in the one-bead-per-residue case.

3.4. Dependence on protein data set of the EDM method

In previous discussions, the same protein data set was used for parameterization and SAXS profile evaluations. This raises concerns about model transferability, as well as the minimum data set size for form-factor optimization. Here, we used different sets of protein structures to perform the parameterization and compared the resulting bead form factors to verify the reliability of our EDM method.

To study the transferability of EDM form factors, the whole data set was split into two subsets containing 503 and 500 structures, respectively, which were randomly assigned to each subset. Two sets of CG form factors were optimized using the respective protein subsets, and the two-bead COE scheme was used in this test. We measured the discrepancy d between the two sets of CG form factors by

Figure 4

Computed SAXS profiles of four example proteins using two-bead-per-residue models. As a reference, the all-atom SAXS profile of each protein is shown as a red dotted line. CG results using the EDM method (COE), the IBA method (COE) and the IBA method ($C\alpha + \text{COM}$) are shown in green, blue and magenta, respectively. The PDB entries of the four example proteins are (a) 4co8, with 98 residues, (b) 4jl5, with 203 residues, (c) 4uyb, with 400 residues, and (d) 1n62, with 804 residues.

$$d = \frac{1}{N_{\text{CG}}} \sum_{\text{CG type}} \left\{ \frac{1}{N_q} \sum_q [f(q) - f_0(q)]^2 \right\}. \quad (17)$$

Here, N_{CG} is the number of CG types, which is 21 for the two-bead COE scheme, and N_q is the number of data points on each CG form-factor curve. It was found that the value of d between the two sets of CG form factors is only 5.49×10^{-3} . As the absolute scale of the CG form factors is about -5 – 5 , this difference corresponds to a relative error of around 1.5%, indicating that the two sets of CG form factors are almost identical.

Additionally, we verified the form factors further by comparing the generated CG SAXS profiles. Two example proteins, lysozyme (PDB code 1lyz; Diamond, 1974) and ubiquitin (PDB code 1ubq; Vijay-Kumar *et al.*, 1987), were used, and the differences in the resulting SAXS profiles are shown in supplementary Fig. S6. It is seen that the discrepancy between the CG SAXS profiles is within the line thickness. The relative differences between the CG SAXS profiles are within 6‰ for lysozyme and 3‰ for ubiquitin. We also

evaluated the R -value distributions for the two sets of CG form factors, using the all-atom SAXS profiles as the standard. In this test, if one protein set was used for parameterization, the other set was then used for SAXS profile calculations. As can be seen in Fig. 6(a), there is only a very small difference between the two distributions. Our test results strongly suggest that the form-factor values do not depend on the protein data set selection. Thus, the demonstrated high accuracy of our EDM method is transferable to various proteins.

To investigate further the effect of protein data set size, we compared the sets of CG form factors obtained from different numbers of proteins using the EDM method, ranging from 100 to 1000. We calculated the deviations from the result with largest data set (1003 structures) for different sizes of protein data set, using the d value defined previously. From Fig. 6(b), the d value decreases from 1.33×10^{-2} to 2.14×10^{-6} as the number of protein structures in the parameterization increases from 100 to 900. However, the greatest reduction in d happens before a data set size of 300, reaching a very small value of 1.98×10^{-3} . This subtle difference suggests that the CG form factors obtained from the EDM method have reached

Figure 5

Computed SAXS profiles of four example proteins using one-bead-per-residue models. As a reference, the all-atom SAXS profile of each protein is shown as a red dotted line. CG results using the EDM method (COE), the IBA method (COE) and the IBA method ($C\alpha$) are shown in green, blue and magenta, respectively. The PDB entries of the four example proteins are (a) 4co8, with 98 residues, (b) 4jl5, with 203 residues, (c) 4uyb, with 400 residues, and (d) 1n62, with 804 residues.

convergence for the size of the protein data set (1003) used in the parameterization.

3.5. Accurate reproduction of experimental scattering profiles

For effective structure modelling, it is necessary to verify that the CG SAXS profiles evaluated from the derived CG form factors are comparable to experimental scattering profiles. There are two solvent effects that affect the comparison significantly. One is from the protein hydration layer. An explicit hydration layer was taken into account, as described in §2.3.1. The other is from the protein displaced solvent. There may be errors when estimating the protein displaced solvent volume using the summation of van der Waals volumes of isolated protein atoms, and the density

reduction of the hydration layer is usually hard to predict. Thus, two free scaling parameters are implemented in many software packages, such as *CRY SOL*, for experimental profile fitting. In our approach, the CG form factor obtained above was decomposed into a vacuum term, representing the contribution of the protein atoms, and an excluded-volume term, representing the contribution of the displaced solvent molecules. The two weights (scaling factors) of the hydration layer and the displaced volume were optimized to obtain reasonable fitting, as described in §2.3.3.

Fig. 7 shows the decomposition of the CG form factor for the backbone CG bead. The original CG form factor (red curve) was optimized using the atomic form factors that consider the excluded-volume correction using the Gaussian sphere approximation, as shown in equation (4). The vacuum term (green curve) was obtained by a separate parameterization using the atomic form factors defined in equation (3), without considering the excluded volumes. Thus, the difference between the two curves is the contribution of the excluded-volume effect (blue curve). This term was then scaled by a weight factor w_{excl} to generate the best fit with experimental scattering profiles.

We then compared the CG SAXS profiles computed using our CG form factors with experimental SAXS profiles for six proteins: lysozyme (PDB code 1lyz; Diamond, 1974), ubiquitin (PDB code 1ubq; Vijay-Kumar *et al.*, 1987), myoglobin (PDB code 1wla; Maurus *et al.*, 1997), RNase (PDB code 1c0b; Bell, 1999), cytochrome *c* (PDB code 1hrc; Bushnell *et al.*, 1990)

Figure 6
Choice of protein structure database for EDM form-factor derivation. (a) Distribution of R values describing the SAXS difference between CG and atomistic methods for each half of the total of 1003 protein structures. Resulting R -value distributions are represented in red and blue, respectively. (b) Convergence of the protein data set size used for the derivation of CG form factors. The data points show the normalized differences between the form factors from a protein data set of N_{PDB} structures and those from the largest protein data set of 1003 structures. The range of N_{PDB} is from 100 to 1003. The definition of the normalized difference can be found in §3.2.

Figure 7
Decomposition of the CG form factor representing the backbone of residues into a vacuum term and an excluded-volume term. The red curve is the fitted CG form factor using the atomic form factors f^{excl} shown in equation (4), which considers the excluded-volume effect. The green curve is the CG form factor evaluated using the original atomic form factors shown in equation (3), without considering the excluded-volume effect. The blue curve is the difference between the red and green curves, which represents the contribution of protein excluded volume. The unit of the form factors is the number of electrons. This term can be scaled to account for the error when evaluating protein excluded volume using the summation of the van der Waals volumes of protein atoms.

and chymotrypsinogen A (PDB code 2cga; Wang *et al.*, 1985). The experimental SAXS profiles of these proteins were obtained from the Small-Angle Scattering Biological Data Bank (SASBDB; Valentini *et al.*, 2015); their entry codes are listed in Table 2. Data points with q values smaller than 0.0311, 0.0601, 0.0248, 0.0631, 0.0201 and 0.0448 $\text{\AA}^{-1}$, respectively, were discarded for the experimental scattering intensities. This is because the Guinier plots of the scattering intensities are nonlinear or the experimental curves are too noisy in this region. We took account of the contribution of the hydration layer and the scaling of the excluded-volume term in the computation of the CG SAXS profiles. The weight for the excess electron density of the hydration layer over bulk water w_{wat} and the weight of the excluded-volume contribution w_{excl} were tuned in order to obtain the best fit to the experimental scattering intensities. We started the optimizations with the default weighting parameters $w_{\text{wat}} = 3\%$ and $w_{\text{excl}} = 1.0$, and the two parameters were tuned using the Nelder–Mead simplex algorithm. It should be noted that similar fitting parameters are available in *CRY SOL* (Svergun *et al.*, 1995). The fitting results for lysozyme and ubiquitin are shown in Figs. 8(a) and 8(b), respectively. Additionally, the fitting results of the other four proteins are shown in supplementary Fig. S7. The results of the fitting parameters and the final χ^2 values from both our method and *CRY SOL* are listed in Table 2.

It is quite clear from Fig. 8(a) that the coarse-grained scattering intensities (blue curve) of lysozyme computed using our CG form factors fit the experimental intensities (red curve) very well. The χ^2 value is only 0.895. This small χ^2 value is very similar to the result from *CRY SOL* ($\chi^2 = 1.404$). During the optimization of the two fitting parameters, a smaller weighting of the hydration shell ($w_{\text{wat}} = 0.864\%$) was included, while a larger excluded volume ($w_{\text{excl}} = 1.051$) was used to obtain the best fit to the experimental scattering profile. With *CRY SOL*, the excess weight of the hydration shell w_{wat} was optimized at 1.5% and the effective atomic radius $r_0 = 0.166$ nm, corresponding to $w_{\text{excl}} = 1.025$ in our formula.

For the ubiquitin results shown in Fig. 8(b), good agreement between the EDM-computed SAXS profile and experimental data is also obtained. Note that the χ^2 value of our CG SAXS profile after tuning the two parameters is 0.960, which is slightly smaller than the value of 1.766 obtained with *CRY SOL*. However, the fitted curves from both methods agree reasonably well with the experimental profile. The value of w_{wat} was optimized at 0.956% and w_{excl} was 1.050, while the parameters optimized by *CRY SOL* were $w_{\text{wat}} = 3\%$ and $w_{\text{excl}} = 1.074$.

Similar χ^2 values were also obtained using our method and *CRY SOL* for the other four example proteins (supplementary Fig. S7). This further confirms the reliability of our CG form factors in modelling experimental SAXS profiles. By comparing the w_{excl} and w_{wat} values from our approach and *CRY SOL*, we find that the two sets of w_{excl} values for the two proteins are highly consistent. This indicates that our excluded-volume term has similar properties to its counterpart in *CRY SOL*. Meanwhile, the w_{wat} values from the two

Table 2

The best-fit weighting parameters yielding optimal χ^2 values for experimental SAXS data from all six example proteins, using either EDM or *CRY SOL*.

The weighting parameter w_{wat} is used to account for the scattering contribution of excess electron density in a surrounding hydration layer, while w_{excl} accounts for the effect of protein excluded volume. The SASBDB entries of the experimental SAXS profiles are listed under the protein name.

Protein	Method	w_{wat} (%)	w_{wat} ($\text{e \AA}^{-3}$)	w_{excl}	χ^2
Lysozyme (SASDAC2)	EDM	0.864	0.003	1.051	0.895
	<i>CRY SOL</i>	1.5	0.005	1.025	1.404
Ubiquitin (SASDAQ2)	EDM	0.956	0.003	1.050	0.960
	<i>CRY SOL</i>	3.0	0.010	1.074	1.766
Myoglobin (SASDAK2)	EDM	0.919	0.003	1.074	0.839
	<i>CRY SOL</i>	2.1	0.007	1.000	0.984
Rnase (SASDAN2)	EDM	2.191	0.007	1.060	1.514
	<i>CRY SOL</i>	5.4	0.018	0.914	1.582
Cyt-C (SASDAB2)	EDM	0.134	0.000	0.952	1.490
	<i>CRY SOL</i>	1.5	0.005	0.975	1.040
Chymotrypsinogen A (SASDAA8)	EDM	1.514	0.005	0.941	2.590
	<i>CRY SOL</i>	3.9	0.013	0.951	2.726

Figure 8

Comparison of EDM-based and experimental SAXS profiles. Two example proteins, (a) lysozyme and (b) ubiquitin, are used. Their experimental SAXS profiles (in red) were obtained from the SASBDB database (<http://www.sasbdb.org/>). The EDM-based computed SAXS profiles are shown in blue. To achieve a best fit between the computed and experimental SAXS profiles, the scattering contributions from the surrounding hydration layer and from the protein excluded volume are optimized.

approaches are quite different, which may originate from the different treatments of the protein hydration layer.

4. Conclusions

We have presented a novel parameterization method to generate coarse-grained form factors for amino acid beads. This method, called electron-density matching, improves the accuracy of the bead form factors by fitting the Fourier transform of the protein electron density calculated from atomic coordinates and form factors. The EDM-based approach can greatly improve the computational efficiency of SAXS intensity modelling, while maintaining an accuracy comparable to methods using atomic form factors. The EDM method is implemented for developing both one-bead- and two-bead-per-residue form factors.

By considering the contribution of the protein hydration layer and the correction of protein excluded volume, the EDM-optimized CG SAXS profiles computed using our CG form factors can precisely reproduce the selected experimental SAXS profiles. Our method is of great potential in *ab initio* structure modelling based on experimental SAXS profiles. Other possible applications include computing SAXS profiles of very large protein complex systems or those from low-resolution structures (e.g. from cryo-electron microscopy). Future work could include applying the same method to other biomacromolecular systems, such as DNA.

All six sets of CG form factors can be found in the supporting information (Tables S1–S6). These form-factor data are also provided in plain text format in a ZIP file, also in the supporting information. Supplementary Table S7 shows the *R* values from the EDM form factors for the six intrinsically disordered proteins. The CG form factors evaluated using the other three different CG schemes are shown in supplementary Figs. S1 and S2. The *R*-value distributions for $q \leq 0.35 \text{ \AA}^{-1}$ are shown in supplementary Figs. S3(a) and S3(b), for two- and one-bead models, respectively. Supplementary Fig. S4 depicts the distribution of *R* values for the six intrinsically disordered proteins, which is based on the EDM method with two-bead COE mapping. In addition, we computed the differences between EDM-based and all-atom SAXS profiles on the basis of a protein data set defined by Stovgaard *et al.* (2010). The corresponding distribution is shown in supplementary Fig. S5, compared with the result from the previous work. The CG SAXS profiles of lysozyme and ubiquitin computed using the two different sets of CG form factors from two different protein data sets are shown in supplementary Fig. S6. Finally, the comparisons of fitted CG SAXS profiles with experimental ones for the four additional protein examples are shown in supplementary Fig. S7.

The program to perform our EDM form-factor parameterization is available from the authors on request.

Acknowledgements

This work was supported by a Tier 1 grant (No. RG45/14) and a Tier 3 grant (MOE2012-T3-1-008) from the Ministry of

Education, Singapore, to LL, and by a National Institute of Health grant (No. R01GM114056) to SY.

References

- Bayer, P., Arndt, A., Metzger, S., Mahajan, R., Melchior, F., Jaenicke, R. & Becker, J. (1998). *J. Mol. Biol.* **280**, 275–286.
- Bell, J. A. (1999). *Protein Sci.* **8**, 2033–2040.
- Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. (1987). *J. Phys. Chem.* **91**, 6269–6271.
- Brown, P. J., Fox, A. G., Maslen, E. N., O’Keefe, M. A. & Willis, B. T. M. (2006). *International Tables for Crystallography*, Vol. C, *Mathematical, Physical and Chemical Tables*, 1st online ed., edited by E. Prince, ch. 6.1, pp. 554–595. Chester: International Union of Crystallography.
- Bushnell, G. W., Louie, G. V. & Brayer, G. D. (1990). *J. Mol. Biol.* **214**, 585–595.
- Chacón, P., Morán, F., Díaz, J. F., Pantos, E. & Andreu, J. M. (1998). *Biophys. J.* **74**, 2760–2775.
- Chen, P. C. & Hub, J. S. (2014). *Biophys. J.* **107**, 435–447.
- Chu, B. & Hsiao, B. S. (2001). *Chem. Rev.* **101**, 1727–1761.
- Cooper, C. D. O., Newman, J. A., Aitkenhead, H., Allerston, C. K. & Gileadi, O. (2015). *J. Biol. Chem.* **290**, 13692–13709.
- Cromer, D. T. & Waber, J. T. (1965). *Acta Cryst.* **18**, 104–109.
- De Guzman, R. N., Martinez-Yamout, M. A., Dyson, H. J. & Wright, P. E. (2004). *J. Biol. Chem.* **279**, 3042–3049.
- Diamond, R. (1974). *J. Mol. Biol.* **82**, 371–391.
- Dobbek, H., Gremer, L., Kiefersauer, R., Huber, R. & Meyer, O. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 15971–15976.
- Doniach, S. (2001). *Chem. Rev.* **101**, 1763–1778.
- Fei, X., Ye, X., LaRonde, N. A. & Lorimer, G. H. (2014). *Proc. Natl Acad. Sci. USA*, **111**, 12775–12780.
- Förster, F., Webb, B., Krukenberg, K. A., Tsuruta, H., Agard, D. A. & Sali, A. (2008). *J. Mol. Biol.* **382**, 1089–1106.
- Franke, D. & Svergun, D. I. (2009). *J. Appl. Cryst.* **42**, 342–346.
- Fraser, R. D. B., MacRae, T. P. & Suzuki, E. (1978). *J. Appl. Cryst.* **11**, 693–694.
- Golub, G. & Kahan, W. (1965). *J. Soc. Ind. Appl. Math. Ser. B Numer. Anal.* **2**, 205–224.
- 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.
- Hura, G. L., Menon, A. L., Hammel, M., Rambo, R. P., Poole, F. L., Tsutakawa, S. E., Jenney, F. E., Classen, S., Frankel, K. A., Hopkins, R. C., Yang, S. J., Scott, J. W., Dillard, B. D., Adams, M. W. W. & Tainer, J. A. (2009). *Nat. Methods*, **6**, 606–612.
- Izvekov, S. & Voth, G. A. (2005). *J. Chem. Phys.* **123**, 134105.
- Jernigan, R. L. & Bahar, I. (1996). *Curr. Opin. Struct. Biol.* **6**, 195–209.
- Kerns, S. J., Agafonov, R. V., Cho, Y. J., Pontiggia, F., Otten, R., Pachov, D. V., Kutter, S., Phung, L. A., Murphy, P. N., Thai, V., Alber, T., Hagan, M. F. & Kern, D. (2015). *Nat. Struct. Mol. Biol.* **22**, 124–131.
- Koch, M. H. J., Vachette, P. & Svergun, D. I. (2003). *Q. Rev. Biophys.* **36**, 147–227.
- Köfinger, J. & Hummer, G. (2013). *Phys. Rev. E*, **87**, 052712.
- Lee, C. W., Martinez-Yamout, M. A., Dyson, H. J. & Wright, P. E. (2010). *Biochemistry*, **49**, 9964–9971.
- Li, Y. Q. & Zhang, Y. (2009). *Proteins*, **76**, 665–676.
- Lipfert, J. & Doniach, S. (2007). *Annu. Rev. Biophys. Biomol. Struct.* **36**, 307–327.
- Liu, H., Morris, R. J., Hexemer, A., Grandison, S. & Zwart, P. H. (2012). *Acta Cryst.* **A68**, 278–285.
- Lyubartsev, A. P. & Laaksonen, A. (1995). *Phys. Rev. E*, **52**, 3730–3737.
- MacKerell, A. D. *et al.* (1998). *J. Phys. Chem. B*, **102**, 3586–3616.
- Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P. & de Vries, A. H. (2007). *J. Phys. Chem. B*, **111**, 7812–7824.

- Maurus, R., Overall, C. M., Bogumil, R., Luo, Y., Mauk, A. G., Smith, M. & Brayer, G. D. (1997). *Biochim. Biophys. Acta (BBA) Protein Struct. Mol. Enzymol.* **1341**, 1–13.
- Mertens, H. D. T. & Svergun, D. I. (2010). *J. Struct. Biol.* **172**, 128–141.
- Miyazawa, S. & Jernigan, R. L. (1985). *Macromolecules*, **18**, 534–552.
- Miyazawa, S. & Jernigan, R. L. (1996). *J. Mol. Biol.* **256**, 623–644.
- Nelder, J. A. & Mead, R. (1965). *Comput. J.* **7**, 308–313.
- 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.
- Noid, W. G., Chu, J. W., Ayton, G. S., Krishna, V., Izvekov, S., Voth, G. A., Das, A. & Andersen, H. C. (2008). *J. Chem. Phys.* **128**, 244114.
- 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.
- Petoukhov, M. V. & Svergun, D. I. (2007). *Curr. Opin. Struct. Biol.* **17**, 562–571.
- Poitevin, F., Orland, H., Doniach, S., Koehl, P. & Delarue, M. (2011). *Nucleic Acids Res.* **39**, W184–W189.
- Ponti, A. (1999). *J. Magn. Reson.* **138**, 288–297.
- 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.
- Reith, D., Pütz, M. & Müller-Plathe, F. (2003). *J. Comput. Chem.* **24**, 1624–1636.
- Rowell, J. P., Simpson, K. L., Stott, K., Watson, M. & Thomas, J. O. (2012). *Structure*, **20**, 2014–2024.
- Šali, A. & Blundell, T. L. (1993). *J. Mol. Biol.* **234**, 779–815.
- 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.
- Schneidman-Duhovny, D., Kim, S. J. & Sali, A. (2012). *BMC Struct. Biol.* **12**, 17.
- Singarapu, K. K., Tonelli, M., Chow, D. C., Frederick, R. O., Westler, W. M. & Markley, J. L. (2011). *J. Biol. Chem.* **286**, 43447–43453.
- Sippl, M. J. (1995). *Curr. Opin. Struct. Biol.* **5**, 229–235.
- Song, J., Lee, M. S., Carlberg, I., Vener, A. V. & Markley, J. L. (2006). *Biochemistry*, **45**, 15633–15643.
- 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. I. (1999). *Biophys. J.* **76**, 2879–2886.
- 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.
- Svergun, D. I., Petoukhov, M. V. & Koch, M. H. J. (2001). *Biophys. J.* **80**, 2946–2953.
- Toft, K. N., Vestergaard, B., Nielsen, S. S., Snakenborg, D., Jeppesen, M. G., Jacobsen, J. K., Arleth, L. & Kutter, J. P. (2008). *Anal. Chem.* **80**, 3648–3654.
- Valentini, E., Kikhney, A. G., Previtali, G., Jeffries, C. M. & Svergun, D. I. (2015). *Nucleic Acids Res.* **43**, D357–D363.
- Vijay-Kumar, S., Bugg, C. E. & Cook, W. J. (1987). *J. Mol. Biol.* **194**, 531–544.
- Wang, D. C., Bode, W. & Huber, R. (1985). *J. Mol. Biol.* **185**, 595–624.
- Wang, G. L. & Dunbrack, R. L. (2003). *Bioinformatics*, **19**, 1589–1591.
- Wassenaar, T. A., Pluhackova, K., Böckmann, R. A., Marrink, S. J. & Tieleman, D. P. (2014). *J. Chem. Theory Comput.* **10**, 676–690.
- Wright, D. W. & Perkins, S. J. (2015). *J. Appl. Cryst.* **48**, 953–961.
- Yang, S. C. (2014). *Adv. Mater.* **26**, 7902–7910.
- 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.