Computer modeling of solution X-ray scattering intensity for biomacromolecules

Tong, Dudu

2016

Tong, D. (2016). Computer modeling of solution X-ray scattering intensity for biomacromolecules. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/69225>

<https://doi.org/10.32657/10356/69225>

NANYANG
TECHNOLOGICAL
UNIVERSITY

COMPUTER MODELING OF SOLUTION X-RAY
SCATTERING INTENSITY FOR BIOMACROMOLECULES

TONG DUDU

SCHOOL OF BIOLOGICAL SCIENCES

2016

COMPUTER MODELING OF SOLUTION X-RAY
SCATTERING INTENSITY FOR BIOMACROMOLECULES

TONG DUDU

SCHOOL OF BIOLOGICAL SCIENCES

A thesis submitted to the Nanyang Technological University

in partial fulfilment of the requirement for the degree of

Doctor of Philosophy

2016

Acknowledgements

Firstly, I would like to express sincere gratitude to my supervisor Asst. Prof. Lu Lanyuan. Your immense knowledge and enthusiasm for research have encouraged me a lot to grow towards a research scientist. Specially, I would like to thank you for your great patience of guiding me throughout my Ph. D. duration. Your advice on both my research works and my future career has been priceless.

My sincere thanks also go to my thesis advisory committee members, Associate Prof. Mu. Yuguang and Asst. Prof. Yang Liang for their insightful comments on my research works. Besides, I would like to thank Associate Prof. Yang Sichun in Case Western Reserve University for the help on my publications.

I would also like to thank other members in our research group: Dr. Nalaparaju Anjaiah, Dr. Chen Wenwen, Mr. Zhu Guanhua, Mr. Bope Christian, Mr. Zhang Haiping and Ms. Wei Kexin for their helpful discussions and valuable suggestions. Mr. Zhang Tong and Ms. Luo Di in the research group of Associate Prof. Mu. Yuguang also help me a lot.

Last but not the least, I would like to express my special thanks to my beloved family for support me spiritually through my Ph. D. life and thesis writing.

Table of Contents

Acknowledgements.....	i
Table of Contents	ii
List of Figures	vi
List of Tables	xvi
Abbreviations	xviii
Abstract	xx
1 Introduction.....	1
1.1 Solution X-ray scattering experiment	1
1.2 Theoretical modeling of solution X-ray scattering intensity.....	4
1.3 Structure modeling from solution X-ray scattering data.....	13
1.4 Outline of the research work	16
2 General methodologies	20
2.1 Atomic form factors and the Debye formula	20
2.2 Spherical average method to compute the orientational average	22
2.3 Evaluation of the scattering intensity based on single structure	24
2.4 Evaluation of the scattering intensity based on MD trajectories.....	25

2.5	Comparison between scattering profiles	27
2.6	Construction of non-redundant protein structure dataset	29
2.7	Protein coarse-grained mapping schemes	30
3	Modeling solution X-ray scattering of biomacromolecules using an explicit solvent model and the fast Fourier transform	32
3.1	FFT algorithm implemented to orientational average	32
3.2	Details of MD simulations and our FFT-based method	36
3.3	Accuracy of the FFT-based method	38
3.4	The computational efficiency of our FFT-based method	41
3.5	Comparison of the scattering profiles with experimental data	45
3.6	The effect of different envelope size	51
3.7	Conclusions	55
4	Accurate optimization of amino acid form factors for computing small-angle X-ray scattering intensity of atomistic protein structures	57
4.1	Methods to derive amino acid form factors	57
4.1.1	Electron density matching (EDM) method	57
4.1.2	Isolated bead approximation (IBA) method	61
4.2	Method to model experimental SAXS profiles	63
4.2.1	Determine the contribution of protein hydration layer	63

4.2.2	Scalable excluded volume.....	66
4.2.3	Optimization of weight factors	67
4.3	Amino acid form factors	67
4.4	Accuracy and efficiency of CG form factors	72
4.5	Example scattering profiles	85
4.6	Dependency of our EDM method to the protein dataset.....	87
4.7	Accurate reproduction of experimental scattering profiles	91
4.8	Conclusions	98
5	Implicit modeling of protein hydration shell in solution small-angle X-ray scattering using modified form factors	99
5.1	Method to parameterize the implicit hydration term.....	99
5.1.1	Computation of target scattering intensities	101
5.1.2	Evaluation of scattering intensity using the implicit hydration term	102
5.1.3	Optimization of the hydration term for atomistic models.....	103
5.1.4	Optimization of the hydration term for coarse-grained models.	105
5.2	Structure optimization using the optimized hydration term.....	106
5.2.1	Pseudo energy function as optimization target	107
5.2.2	Iterative optimization based on analytical derivatives	110

5.3	Resulting implicit hydration term	112
5.4	Accuracy of SAXS profile evaluation.....	114
5.5	Robustness of the fitting process.....	123
5.6	Result of structure optimization	127
5.7	Conclusions	132
6	Concluding remarks	134
	References.....	136
	Publications.....	147

List of Figures

- Figure 1.1 Illustration of a solution X-ray scattering experiment. X-rays with a certain wavelength are focused by a collimator. After passing through the sample solution, the scattering intensity at different scattering angles are measured by the detector. The resulting scattering intensity data are obtained by the difference between the intensity of sample solution and pure solvent. Figure is adopted from ref.[14].2
- Figure 1.2 A typical scattering intensity curve of Myoglobin. The curve at different range of q contains different types of information. The molecular weight, rough shape and radius of gyration (R_g) of the protein are encoded in the low q region. Moreover, the medium range of the q curve corresponds to the protein fold and secondary structure. Finally, the high q region contains the atomistic details of the protein structure, though the information is orientationally averaged. The scattering intensity data to generate the plot is obtained from Small angle scattering biological databank (SASBDB) [15].3
- Figure 2.1 Atomic form factors for typical atoms that exist in protein molecules. The solid lines represent the original form factors, as computed in equation (2.1). And the dashed lines show corresponding form factors with the excluded volume correction, shown in equation (2.2).21
- Figure 3.1 (a) the R values between the scattering intensities computed from the Debye formula and our FFT-based method when different $q_{\max}$ values are used. (b) Overlapping of the scattering profiles computed from the Debye formula, the spherical average method and our FFT-based approach.40

Figure 3.2 The comparison of computational times consumed by the spherical average method and our FFT-based method for the three benchmark proteins. The computational times are estimated using 101 frames of MD simulation trajectories. Data points on each curve correspond to different hydration shell thickness used to construct the envelope. They are 0.5, 0.7, 1.0, 1.5, 2.0, 2.5 and 3.0 nm respectively. The results of Lysozyme, Myoglobin and RNase are colored in red, green and blue respectively. The spherical points represent the computational times of the spherical average method, while the cube points represent those of our FFT-based method. 42

Figure 3.3 The scattering intensities computed using the atomistic MD simulation trajectories by the WAXSiS server and our FFT-based method, compared with the experimental scattering profiles after minimizing the χ^2 values for the three benchmark proteins.47

Figure 3.4 The scattering intensities of Lysozyme computed using the FFT implementation of the AXES approach and the original AXES server, compared with the experimental data after minimizing the χ^2 values.50

Figure 3.5 (a) The scattering intensities of Lysozyme computed using different solvation shell thickness ranging from 0.5 to 3.0 nm. (b) The comparison of the scattering intensities from different solvation shell thickness with experimental scattering intensities.53

Figure 4.1 Correction of CG form factors obtained by the IBA method. The red curve is the original CG form factors of CG bead representing alanine side chain. The positive values at low q region are incorrect. So we use only the values after the reflection point, shown in blue stars, to perform a six-order polynomial fitting in order to obtain reasonable values for the CG form

factors at low q region. The value of $F^{CG}(q = 0)$ evaluated from the summation of all the atomic form factors in the CG bead, shown by the green sphere, is used as fitting constraint. The resulting new CG form factors are shown in green curve.63

Figure 4.2 The coarse-grained scheme and CG form factors obtained from electron density matching (EDM) and single bead approximation (IBA), respectively. (a) Illustration of a two-bead-per-residue CG scheme where each bead is placed at its center of electron (COE). For the illustrative structure shown in sticks, the backbone and side-chain of each amino acid is represented by a blue and red bead, respectively. (b) and (c) Resulting CG form factors from EDM and IBA, respectively, where the blue and red curves represent form factors of backbone and side-chain beads of all 20 residues, respectively. The ordering of their side-chain form factors is following 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. ...68

Figure 4.3 The CG form factors obtained by (b) EDM and (c) IBA, using the one-bead COE CG scheme illustrated in (a). The red curves show the CG form factors representing the 20 types of protein residues. The ordering of the residue form factors is the 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.69

Figure 4.4 The CG form factors obtained by the EDM method using (a) the two-bead $C\alpha + COM$ and (b) one-bead $C\alpha$ CG schemes respectively. The blue curve in (c) shows the CG form factor of the residue backbone. The red curves in (c) represent the CG form factors of the 20 types of residue

side-chains, while the red curves in (d) represent the CG form factors of the 20 types of protein residues. The ordering of the side-chain form factors in (c) and the residue form factors in (d) is the 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.71

Figure 4.5 Distribution of R values describing the difference of SAXS profiles between CG and all-atom models. The SAXS profiles within q range of $0 \sim 0.5 \text{ \AA}^{-1}$ are used in the computation. Two different CG schemes are shown: (a) two-bead-per-residue model and (b) one-bead-per-residue model. The results of EDM and IBA are presented in red and blue curves, respectively, where spherical circles represent that the position of each CG bead is placed at its corresponding COE and cubic points represent 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 center of mass (COM) of its side-chain (in the case of a two-bead-per-residue model).74

Figure 4.6 Distribution of R values describing the difference of SAXS profiles between CG and all-atom models. Two different CG schemes are shown: (a) two-bead-per-residue model and (b) one-bead-per-residue model. The results of EDM and IBA are presented in red and blue curves, respectively, where spherical circles represent that the position of each CG bead is placed at its corresponding COE and cubic points represent 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 center of mass (COM) of its side-chain (in the case of a

two-bead-per-residue model). The R values are computed for the q range of $0 \sim 0.35 \text{ \AA}^{-1}$75

Figure 4.7 Average R values between CG and all-atom SAXS profiles as a function of q_{cut} values. The R values were evaluated over the q -range below q_{cut} and averaged over all the 1003 protein structures. The EDM method was used for the derivation of the CG form factors. Results from both one-bead and two-bead COE CG schemes are shown by blue and red curves, respectively.77

Figure 4.8 Distribution of R values describing the difference of SAXS profiles between CG and all-atom models for a dataset containing NMR structure ensembles of intrinsic disordered proteins. Six proteins (PDB entries: 1A5R, 1R8U, 2FFT, 2L14, 2LJL, and 2LY4) with total 100 structure models are contained in the dataset. The form factors obtained by our EDM method together with the two-bead COE CG scheme are used to compute the CG SAXS profiles.79

Figure 4.9 Comparison of the S value distribution of the EDM method with that reported by Stovgaard et al. [42]. S is evaluated by

$$S = \sqrt{\frac{1}{N_q - 1} \sum_q \left[\frac{I_{CG}(q) - I_{AA}(q)}{\sigma(q)} \right]^2}, \text{ where the empirical standard error}$$

$\sigma(q_i) = I(q_i)(q_i + \alpha)\beta$ with $\alpha=0.15$ and $\beta=0.3$. We use the same protein dataset from Stovgaard et al. The red histogram shows the distribution of S values from our method, and the blue histogram shows the results reported by Stovgaard et al.81

Figure 4.10 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 in a red dotted line. CG results using the EDM method (COE), the IBA method (COE) and the IBA method ($C\alpha$ + 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..84

Figure 4.11 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 in 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..85

Figure 4.12 Comparison of CG SAXS profiles computed from different sets of CG form factors. The red line with spheres shows the CG SAXS profiles evaluated with the CG form factors obtained from protein set 1, while the green line shows the results when protein set 2 is used for parameterization. The definition of the two protein sets can be found in section 3.4 of the main text. The blues lines show the relative errors using the atomistic results as the standard. The two proteins are (a) Lysozyme (1LYZ) with 129 residues and (b) Ubiquitin (1UBQ) with 76 residues.88

Figure 4.13 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 dataset size used for the derivation of CG form factors. The data points show the normalized differences between the form factors from a protein dataset of N_{PDB} structures and those from the largest protein dataset of 1003 structures. The range of N_{PDB} is from 100 to 1003. The definition of the normalized difference can be found in main text.....89

Figure 4.14 Decomposition of the CG form factor representing the backbone of residues to a vacuum term and an excluded volume term. The red curve is the fitted CG form factor using atomic form factors f^{excl} shown in Eq. (2.2), which considers the excluded volume effect. The green curve is the CG form factor evaluated by using the original atomic form factors shown in Eq. (2.1), 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 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.....92

Figure 4.15 Comparison of EDM-based and experimental SAXS profiles. Two example proteins, (a) lysozyme and (b) ubiquitin, are used where their experimental SAXS profiles (in red) are 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 computed and experimental SAXS profiles, the scattering contributions from the surrounding hydration layer and from the protein excluded volume are optimized.....95

Figure 4.16 Comparison of EDM-based and experimental SAXS profiles for four additional example proteins: (a) Myoglobin, (b) RNase, (c) Cytochrome C and (d) Chymotrypsinogen A. The experimental SAXS profiles (in red) are 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 computed and experimental SAXS profiles, the scattering contributions from the surrounding hydration layer and from the protein excluded volume are optimized.96

Figure 5.1 Comparison of the implicit hydration term $f^s(q)$ optimized using atomistic and Ca -based coarse-grained protein structures. The result of all-atom representation is shown in red curve, while the Ca based coarse-grained scheme is shown in blue curve. 113

Figure 5.2 The R value distributions of scattering intensities computed from our dataset containing 880 protein structures. The R values are computed using the scattering intensities evaluated by atomistic protein structures plus explicit hydration layers as targets. (a) Atomistic structure representations are used to compute the scattering intensities. The red and blue curves show the distributions of our implicit hydration treatment and Crysol respectively. (b) Ca -based CG structure representations are used to compute the scattering intensities. The red and blue curves show the distributions of our implicit hydration treatment and Fast-SAXS-pro respectively. 116

Figure 5.3 The comparison of SAXS profiles computed using explicit hydration water molecules and our implicit hydration term, as well as Crysol, for four selected proteins with different sizes. The red spheres are the all-atom

SAXS profiles computed using explicit hydration water molecules. The green lines represent the all-atom SAXS profiles evaluated with the atomic form factors modified by our implicit hydration term and the blue curves illustrate the results computed from Crysol. The four example proteins shown in the figures are: (a) 4CO8 (98 residues), (b) 4JL5 (203 residues), (c) 4UYB (400 residues) and (d) 1N62 (804 residues)..... 118

Figure 5.4 The comparison of SAXS profiles computed using explicit hydration water molecules and our implicit hydration term, as well as Fast-SAXS-pro for four selected proteins with different sizes. The red spheres are the all-atom SAXS profiles computed using explicit hydration water molecules. The green lines represent the coarse-grained SAXS profiles evaluated with the residue-based coarse-grained form factors modified by our implicit hydration term and the blue lines illustrate the results of Fast-SAXS-pro. The four example proteins shown in the figures are: (a) 4CO8 (98 residues), (b) 4JL5 (203 residues), (c) 4UYB (400 residues) and (d) 1N62 (804 residues). 119

Figure 5.5 (a) the optimized implicit hydration term using different numbers of protein structures in our dataset. The result of using the whole dataset (880 structures) is shown in red curve, as the comparison target for computing relative differences. The results of using 200, 400, 600 protein structures to perform optimization are shown in blue, magenta and cyan respectively. (b) The relative differences of the resulting implicit hydration terms computed using different numbers of protein structures from 100 to 800. The comparison target is the red curve shown in (a). The formula to compute the relative difference can be found in the main text..... 124

Figure 5.6 Comparison of implicit hydration terms optimized when different hydration weight is used to calculate the target scattering profiles. For the comparison purpose, the resulting implicit hydration terms are divided by their corresponding hydration weight values respectively.....	125
Figure 5.7 Results of structure optimization using our implicit hydration layer to compute scattering intensity for four example proteins. The initial and target structures are shown in white and red respectively. Our optimization result is shown in green. The atomistic structure of our optimization result used to generate the cartoon representation is recovered from the $C\alpha$ coordinates by the software PULCHRA. The four example proteins are (a) adenylate kinase, (b) LAO binding protein, (c) maltodextrin-binding protein and (d) elongation factor 2.....	130
Figure 5.8 The evolution of the total optimization energy (shown in red curve), SAXS-fitting energy (shown in green curve) and RMSD value (shown in blue curve) during our structure optimization process using LAO binding protein as an example. The jump in the total optimization energy means the starting of a new optimization cycle with a larger λ value, where a larger weight is assigned to the SAXS-fitting energy W_q	131

List of Tables

Table 1.1 Most widely used methods to model theoretical scattering intensities and their features.....	10
Table 3.1 The χ^2 values of the computed scattering profiles calculated by the Waxisis server and our FFT-based method, compared with the experimental scattering intensities of the three benchmark proteins.	46
Table 3.2 The χ^2 values between the computed and experimental scattering intensities for the three benchmark proteins using different solvation shell thicknesses. The protein backbone restraint MD simulation trajectories are used to compute scattering intensities.....	49
Table 3.3 The χ^2 values between the computed and experimental scattering intensities for the three benchmark proteins using different solvation shell thicknesses. The MD simulation trajectories without protein backbone restraint are used to compute scattering intensities.....	52
Table 4.1 Comparison of the mean and standard deviation of R values for the six intrinsic disordered protein NMR structure ensembles, using our EDM method. Results from one and two bead COE models are shown.	79
Table 4.2 The best-fit weighting parameters yielding optimal χ^2 values for experimental SAXS data of all the six example proteins, using EDM and CRY SOL, respectively. The weighting parameter w_{wat} is to account for the scattering contribution of excess electron density in a surrounding hydration layer, while w_{excl} is for the effect of protein excluded volume.	

The SASBDB entries of the experimental SAXS profiles are listed under the protein name.....	94
Table 5.1 Fitting of experimental SAXS profiles using our implicit hydration treatment (with both an atomistic and a C α based coarse-grained structure representations) and Crysol for six selected proteins. The experimental SAXS profiles are obtained from SASBDB, with the entry of each protein listed in the table. The weights of hydration contribution and protein excluded volume are tuned to best reproduce experimental data.	121
Table 5.2 Results of structure optimizations for the four benchmark proteins: adenylate kinase, LAO binding protein, maltodextrin-binding protein and elongation factor 2. For adenylate kinase, both the final RMSD and the best RMSD value obtained during the optimization process are shown.	128

Abbreviations

SXS	Solution X-ray scattering
SAXS	Small angle X-ray scattering
WAXS	Wide angle X-ray scattering
NMR	Nuclear magnetic resonance
PDB	Protein databank
MD	Molecular dynamics
R _g	Radius of gyration
CG	Coarse-grained
COM	Center of mass
COE	Center of electrons
FFT	Fast Fourier transform
EDM	Electron density matching
IBA	Isolated bead approximation
RMSD	Root mean square deviation
RDF	Radial distribution function
SASA	Solvent accessible surface area

IDP Intrinsic disordered protein

SASBDB Small angle scattering biological data bank

Abstract

Increasing amounts of scattering data are obtained from high-throughput Solution X-ray scattering (SXS) experiments, including small-angle X-ray scattering (SAXS) and wide-angle X-ray scattering (WAXS). There is a great demand of computational methods that can retrieve useful structural information or build structure model from these data. We have proposed substantial improvements to current methods that model the scattering profiles from protein structure at both atomistic scale and coarse-grained (CG) scale. In addition, our coarse-grained approach can be conveniently applied to structure optimization based on target scattering intensity.

Firstly, a fast Fourier transform (FFT) based orientational average method is proposed to improve the computational efficiency of modeling scattering profiles using an atomistic protein structure representation, especially in case of considering explicit hydration water molecules. Comparing with the popular spherical average method, our method will become more efficient for systems with more than 3000 atoms. Moreover, the computational time of our FFT-based method remains nearly unchanged as the system size increases, making it suitable for very large protein complexes.

CG representations are also widely used to improve the computational efficiency of theoretical scattering intensity computation. Given the importance of accuracy for CG approaches, we have proposed the electron density matching (EDM) method to parameterize the CG form factors. Comparing with the CG form factors used in literature, our EDM-derived ones result in better agreement to atomistic scattering intensities. Furthermore, the resulting CG

form factors are shown to reproduce the experimental scattering profiles well by including the contribution of hydration layer and the correction of protein excluded volume.

Finally, in order to perform structure modeling with our EDM-derived CG form factors, we have proposed an implicit hydration term to take account the contribution of the hydration layer scattering. This term is only related to the surface accessible solvent area (SASA) of protein atoms, making our formulation to evaluate scattering intensity analytically differentiable to the protein coordinates. The implicit hydration term is fitted to best reproduce the overall scattering intensity computed using explicit hydration water molecules. It is shown that the conjugate gradient structure optimization based on the target scattering intensity can produce final molecular structures very close to the known target structure.

1 Introduction

1.1 Solution X-ray scattering experiment

Solution X-ray scattering (SXS) is an increasingly popular method to study the flexible biological macromolecules structures and complexes [1-13], combined with other conventional structural biology techniques. The solution scattering experimental is schematically illustrated in Figure 1.1. The purified biomacromolecular sample in solution state is exposed to an X-ray beam line, and the scattering intensity $I(q)$ is recorded by the detector as a function of the momentum transfer vector q . The momentum transfer vector is related to the scattering angle 2θ by $q=4\pi\sin\theta/\lambda$, where λ is the wavelength of the input X-rays, typically around 0.15 nm. In SXS experiments, both the scattering intensity of sample solution and pure solvent are collected. The latter intensity is then subtracted from the former, which is a process usually termed as buffer subtraction, in order to remove the background scattering as well as various kinds of errors during the experiment.

Figure 1.1 Illustration of a solution X-ray scattering experiment. X-rays with a certain wavelength are focused by a collimator. After passing through the sample solution, the scattering intensity at different scattering angles are measured by the detector. The resulting scattering intensity data are obtained by the difference between the intensity of sample solution and pure solvent. Figure is adopted from ref.[14].

A typical scattering intensity curve is shown in Figure 1.2. The x-axis is the momentum transfer vector q and the y-axis is the scattering intensity $I(q)$. The intensity curve at different q region corresponds to different types of information [13]. For example, at low q range of $0\sim3\text{ nm}^{-1}$, the intensity decreases rapidly with q to the power of four. The curve mainly contains the information of molecular weight, shape and radius of gyration (R_g). A medium q range of $3\sim10\text{ nm}^{-1}$ contains information of overall fold and secondary structures. The curve at high q region mainly possesses the information of the atomistic structure of the macromolecule. Based on the range of scattering intensity data obtained, SXS experiments are further divided into small-angle X-ray scattering (SAXS) and wide-angle X-ray scattering (WAXS) experiments. Traditional SAXS experiments can accurately measure the scattering intensity till the q range of about 5 nm^{-1} . In recent years, the new development of light sources and detection devices allowed experimentalists to obtain the scattering

intensity to higher q regions with less error, making it possible to model higher-resolution structure details.

Figure 1.2 A typical scattering intensity curve of Myoglobin. The curve at different range of q contains different types of information. The molecular weight, rough shape and radius of gyration (R_g) of the protein are encoded in the low q region. Moreover, the medium range of the q curve corresponds to the protein fold and secondary structure. Finally, the high q region contains the atomistic details of the protein structure, though the information is orientationally averaged. The scattering intensity data to generate the plot is obtained from Small angle scattering biological databank (SASBDB) [15].

Compared with the traditional X-ray crystallography and nuclear magnetic resonance (NMR) experiment to study the structure of biomacromolecules, SXS has the advantage of high throughput due to the simplicity of sample preparation. Neither crystallization nor isotopic labeling is needed before the scattering experiment. In addition, similar to NMR experiment, the structural information obtained is in the solution state, which is very close to the biologically functional state. Both SXS and NMR can also measure a dynamic

ensemble of conformations rather than a single conformation. As a tradeoff, the result of SXS experiment is a one-dimensional curve, in which the information contained is highly degenerated. This is because the positions and orientations of the scattering particles are randomly distributed as they can move and rotate freely in the solution state, supposing that we have a dilute biomacromolecule solution with relatively weak inter-particle interactions. Thus, it is impossible to directly solving the three-dimensional structure from the resulting scattering intensity as carried out in X-ray crystallography and computational tools are needed to help build structure models from experimental scattering profiles.

1.2 Theoretical modeling of solution X-ray scattering intensity

From the fundamentals of physics it is known that X-rays mainly interact with the electrons. So the X-ray scattering profiles of biomacromolecules in vacuum can be computed from the Fourier transforms of their electron density distributions. In practice, because discrete atomic coordinates are easily obtained, the atomic form factors, which are the scattering amplitudes of isolated atoms, are usually used to evaluate the scattering profiles instead of the electron density distributions. The Debye formula [16], which is described in the Method section, is often used to perform such kind of computations.

However, modeling the scattering profiles of biomacromolecules in solution state is not so straightforward. Not only the scattering of the protein molecules, but also the interactions between the protein and solvent molecules need to be considered. The structure of water molecules inside the hydration layer around the protein molecule is reported to be very different from the bulk solvent, due to the interaction with the protein molecule and its density is also slightly

higher [17-19]. Additionally, the buffer subtraction process as in the SXS experiment must also be accounted, and the excluded volume effect of the protein molecules should also be incorporated. There are already many computational tools in literature to evaluate the theoretical scattering intensities given an input structure [20-35]. The major differences between these methods are in the way they treat the contribution of hydration layer and the excluded volume effect.

The most widely used method to model the scattering profile is Crysol, developed by Svergun et al. [20]. Crysol computes the scattering intensity based on a single input protein structure. The excluded volume effect is taken into consideration using the Gaussian sphere approximation proposed by Fraser et al. [36]. In this approximation, it is supposed that the excluded volume effect can be simulated by assuming that the volume of the protein atoms is filled with electron gas. The average electron density should be equal to that of the bulk solvent, and the distribution of the electrons is assumed to be a Gaussian sphere. Then the atomic form factors of the protein atoms are corrected by the Fourier transform of this Gaussian sphere. Additionally, the contribution of hydration solvent molecules is implicitly described by a layer of 3 Å with uniform electron density. There are two free parameters, corresponding to the excluded volume and the electron density of the hydration layer, which can be tuned in order to fit the experimental scattering profile. Spherical harmonics [37] is introduced in Crysol to accelerate the computation of scattering intensity. Similar to Crysol, there are also other methods using implicit treatment of the hydration water [21-28]. For example, AquaSAXS is a web server to compute SAXS profiles using non-uniform electron densities for the hydration layer [28]. The equilibrium solvent density maps around the protein structures are

evaluated based on the physical nature of the protein using methods such as AquaSol [38] or 3D-RISM [39]. Then excess hydration densities are added based on this equilibrium solvent density map, which shows improved performance compared with the uniform excess hydration densities used in Crysol. Another similar tool to implicitly treat the contribution of solvent is FoXS [21, 22], in which the form factors of hydration water molecules are directly incorporated to nearby protein atoms according to their solvent accessible surface areas (SASA). The computational efficiency is improved with the tradeoff of slightly lower accuracy. In general, these methods usually target at the scattering profiles at the SAXS region (up to $q = 5 \text{ nm}^{-1}$) only, as the atomistic structure and the distribution of water are not accurately considered by the implicit treatment.

In contrast to the implicit representation of the hydration layer, Fast-SAXS-pro developed by Yang et al. uses an explicit representation of the solvent layer [29, 30]. The positions of the explicit water molecules are obtained by overlapping the input protein structure with a pre-equilibrated water box and finding out all water molecules within a certain cutoff to the protein surface. In Fast-SAXS-pro, the excluded volume is still evaluated using the Gaussian sphere approximation. Because of the coarse-grained structure representation for both protein and hydration water molecules adopted in the method, it can only fit to the experimental data in the SAXS region. Furthermore, the AXES method proposed by Grishaev et al. [31] considers the excluded volume effect by the explicit solvent molecules in addition to the hydration layer. By using explicit solvent molecules to compute the scattering contribution of protein excluded volume and excess hydration effect, the AXES method can accurately reproduce the experimental scattering profiles until the near-WAXS region ($q =$

10 nm^{-1}). The underlying problem of the above two approaches is that the position of the explicit water molecules are taken from a simulation of bulk water instead of a protein solution simulation. Thus the structure difference between water molecules in the hydration layer and the bulk is neglected.

Different from the AXES method discussed above, there are other methods using molecular dynamic (MD) simulations to consider the “real” structure of hydration water molecules. Park et al. proposed a formula that can be used to compute the scattering intensity using the simulation trajectories of both the protein solution and the pure solvent [32]. Basically, an envelope containing the protein molecule and the hydration water molecules is constructed in the simulation trajectory of the protein solution. Another image envelope with exactly the same shape is constructed in the pure solvent trajectory. The scattering amplitudes of systems inside the two envelopes are used to evaluate the scattering intensity. Chen & Hub use an equivalent formula to build a web server named WAXSiS [33], which can automatically perform MD simulations and scattering intensity computations given an input protein structure. Compared with the original method, the fluctuations of the protein conformation are taken into consideration during the MD simulation process, which results in improved fitting to experimental scattering intensities. Additionally, a similar formula is proposed by Oroguchi et al. also to compute the scattering intensity from MD simulation trajectories [34]. On the other hand, Kofinger & Hummer developed a histogram-based method to model the scattering profiles from explicit solvent simulations [35]. It is shown that by using the MD simulation trajectories, these methods can reproduce experimental results up to the near-WAXS region ($q = 10 \text{ nm}^{-1}$).

Besides the different treatments of the hydration layer and the excluded volume effect, there are also different algorithms to perform the orientational average. As a result of the free movement and rotation of protein molecules in the solution, the experimentally obtained solution scattering profiles are one-dimensional curves of the momentum transfer vector q . The orientational average in the theoretical modeling of scattering intensities is the process to project the three-dimensional scattering intensity at $\mathbf{q}$ space to the one-dimensional profile. The rotation of protein coordinate $\mathbf{r}$ is equivalent to the rotation of the scattering vector $\mathbf{q}$. Thus, the scattering intensities obtained at different scattering vector $\mathbf{q}$ with the same scalar value q need to be averaged in order to obtain the final scattering intensity. Theoretically, the Debye formula can be used to accurately obtain the result of the orientational average. However, the computational time of the Debye formula scales with the number of atoms in the system by $O(N^2)$. For relatively large systems, this can be extremely computational expensive. So it is only used in some coarse-grained approaches in literature, where the number of atoms tends to be small. A simple improvement of the Debye formula is to perform the numerical average over selected directions of $\mathbf{q}$ on the reciprocal q sphere, named as the Spherical Average method. The computational complexity of the Spherical Average method is $JO(N)$, where the pre-factor J is the number of directions selected. Generally, the Spherical Average method will be much fast than the Debye formula, as J is usually smaller than N for even small proteins. The Spherical Average method is widely used in methods such as AXES [31] and WAXSiS [33]. Additionally, the orientational average can be also facilitated by the multipole expansion method using spherical harmonic basis functions. The computational time of this method, as well as the accuracy, depends on how many basis functions are used to perform the approximation. Usually for

irregularly shaped (non-spherical) proteins, higher orders of expansions are needed in order to accurately approximate the scattering intensity. The multipole expansion method is used by Crysol [20] and also Oroguchi et al. [34] to improve the computation efficiency. A similar Zernike polynomial expansion is used by Liu et al. [26]. The histogram based method used by Kofinger & Hummer [35] first evaluates the radial distribution function (RDF) between different atom pairs, and then performs Fourier transform to the RDF to calculate the scattering profile. The method still has the computational complexity of $O(N^2)$ during the computation of RDF. However, the pre-factor is quite small as there is no need to evaluate sin/cos functions. So for certain system sizes, the method is still quite efficient. We have proposed a fast Fourier transform (FFT) based method to improve the computational efficiency of the orientational processing [40]. The computational time of our method remains nearly unchanged as the system size increases, thus it is very useful when evaluating the scattering intensities of very large systems. Table 1.1 summarizes those popular methods to evaluate the scattering intensities in literature, as well as their features.

Table 1.1 Most widely used methods to model theoretical scattering intensities and their features.

Methods	Representation	Excluded volume	Hydration layer	Orientational average
Crysol	Atomic	Gaussian sphere	Protein shape function	Multipole expansion
AquaSAXS	Atomic	Gaussian sphere	AquaSol solvent density distribution	Cubature formula
FoXS	Atomic or residual level CG	Gaussian sphere	Implicit treatment based on SASA	Debye formula
Stovgaard et al.	CG (1 or 2 bead per residue)	CG form factor	-	Debye formula
Zernike expansion	Atomic	Uniform density	Voxelized representation	Zernike expansion
Fast-SAXS-pro	Residual level CG	Gaussian sphere	Explicit CG water	Debye formula
AXES	Atomic	Explicit water	Explicit water	Spherical average
Park et al.	Atomic	Explicit water (MD)	Explicit water (MD)	Spherical average
WAXSiS	Atomic	Explicit water (MD)	Explicit water (MD)	Spherical average
Oroguchi et al.	Atomic	Explicit water (MD)	Explicit water (MD)	Multipole expansion
Kofinger & Hummer	Atomic	Theoretical water RDF	Explicit water (MD)	RDF based

In general, SAXS profile contains coarse-grained structural information, as the resolution of technique is relatively low. Thus, in order to further improve the computational efficiency, coarse-grained (CG) protein structure representations are used by some methods in literature [25, 29, 30, 41, 42]. Those methods differ at the CG schemes used and the approach to obtain CG form factors. For example, Fast-SAXS-pro [29, 30] uses a one-bead per-residue CG scheme, where the position of the CG bead is placed at the C α position of the corresponding residue. The coarse-grained form factors used in Fast-SAXS-pro are simply computed using the Debye summation over all the atoms contained in each coarse-grained group, taking average over all possible conformations observed in a selected non-redundant protein structure dataset. This method is termed as Isolated Bead Approximation (IBA) here. Notably, such method is also referred as Single Bead Approximation (SBA) by Niebling et al. [41] in their later work. Zheng et al. [25] has adopted a similar one-bead per-residue approach. Different from Fast-SAXS-pro, the hydration water molecules are implicitly incorporated into coarse-grained beads located in the protein surface. This further improved the computational efficiency by reducing the hydration water degrees of freedom. The MARTINI coarse-grained scheme, which generally maps four heavy atoms to one coarse-grained bead, is used by Niebling et al very recently [41]. This relatively finer coarse-grained scheme results in more accurate CG SAXS profiles, using profiles of atomistic ones as the standard. However, no hydration is considered in this work, making the comparison with real experimental data difficult. As an alternative, the result is used to reproduce the difference scattering data from experiment, where the solvent contributions are canceled out. The IBA or IBA-like approaches mention above do not consider the correlations between different coarse-grained beads inside the proteins, as each bead is regarded as an isolated

bead only. Thus the accuracy of the CG form factors is limited, especially for the relatively low resolution coarse-grained schemes. This inspires the development of more accurate coarse-grained form factors. In the work of Niebling et al, the author also mentioned a self-consistent approach to consider the correlations between coarse-grained beads and obtained a set of CG form factors that can reproduce all the pairwise bead-bead scattering intensities. Nevertheless, the numerical convergence is not guaranteed for the algorithm in this approach. Additionally, Stovgaard et al. [42] proposed to use the Markov chain Monte Carlo method based on the maximum-likelihood to parameterize CG form factors for both one-bead per-residue and two-bead per residue schemes. The fitting targets are chosen to be the theoretical SAXS profiles computed using Crysol without considering the contribution of the protein hydration shell. Although the coarse-grained SAXS profiles computed using the obtained CG form factors overlapped with the theoretical targets very well within the q range of $0 \sim 0.75 \text{ \AA}^{-1}$ for the two-bead per-residue coarse-grained mapping scheme, the result is not directly comparable with the experimental SAXS profiles as the hydration layer is not included. In our work, the electron density matching (EDM) method is proposed to parameterize the CG form factors based on the least square fitting of the Fourier transform of the protein electron density distributions. The resulting CG form factors reproduce the atomistic scattering intensities very well, especially for relatively higher resolution CG representations.

1.3 Structure modeling from solution X-ray scattering data

In the field of solution X-ray scattering there are mainly two problems to address and they are reversible to each other. The first one is given a structure model, how to evaluate the scattering intensities of this model. As discussed in the previous part, there are already many methods in literature to perform the computation. The reverse one, and also the more challenging problem, is given an experimental solution X-ray scattering profile, how to build the corresponding structure model. The one-dimensional scattering profile does not contain enough information to fully determine the three-dimensional structure of the protein. Thus, the common procedure to perform structure modeling usually starts from an initial guess. The initial model's theoretical scattering intensity is computed and compared with the target experimental one. Then the structure model is iteratively modified until the computed scattering profile overlaps with the experimental one reasonably well. It is seen that the structure modeling based on experimental scattering profiles needs not only accurate evaluation of scattering intensities from certain structure models but also reasonable sampling of the conformational space. This makes the structure modeling much more difficult than evaluating scattering intensity from a given structure. There are already some methods that can perform such kind of modeling in literature [25, 43-46]. However, the structure modeling still remains an open problem in the community. Until very recent papers, the success rate of the high resolution structure modeling remains relatively low [46]. Svergun proposed a method to restore the rough shape and low resolution structure of protein from the solution scattering data. The protein is represented by a set of dummy atoms enclosed in a sufficiently large volume. Simulated annealing algorithm is then performed to determine which dummy atoms

belong to the protein based on the fitting of the target scattering profile. This method is implemented in the program DAMMIN [43]. Later, Franke et al. improved the method by optimizing the speed and accuracy as well as removing the artifacts during the conformational searching process. The improved method is implemented as DAMMIF [44]. DAMMIN and DAMMIF are quite powerful as *ab initio* modeling of the protein low resolution structure is performed, without any prior knowledge of the protein structure. However, one of the problems for these two methods is no physical potential is applied to restrain the sampling of structure, partially due to that the dummy atoms are not related to real protein atoms or residues. Thus it is possible that the optimization results are unreasonable structures. Similar to the above two methods, GASBOR [47] is another method to perform *ab initio* structure modeling from SAXS data. The protein structures are represented by chain-like ensemble of dummy residues. However, there are still no physical potentials to describe the interactions between residues and the use of averaged residue form factors has greatly lowered the accuracy. By contrast, Zheng et al. [25] used a residue-based coarse-grained representation to perform structure optimization based on target scattering intensities. Besides the scoring function to measure the similarity of scattering intensity, a modified elastic network model potential is also applied to constrain the protein structure during the optimization process. The modified elastic network model is constructed from a known crystal structure of the protein, which serves as the input. And the structure is biased to this input crystal structure. Thus the method requires a known initial structure that is not far away from the real solution. The derivative of the scoring function containing the modified elastic network model and the SAXS-fitting score is computed numerically and the Newton-Raphson algorithm is used during the structure optimization process. Gorba et al. [45] , also employing an elastic

network model, developed an iterative normal mode analysis method to fit the pair distribution function derived from the experimental scattering intensity data. The linear combination of low-frequency mode is used to generate small structure deformations during each iteration step. This method also needs the input of a high-resolution initial structure of the target protein. Very recently, Panjkovich et al. [46] presented a hybrid method termed SREFLEX to perform the structure modeling. Hierarchical optimizations containing two steps are included in the method. The first step is to probe large domain movements, while the second step focuses on smaller and more localized displacements. The conformational changes in both steps are obtained by the normal mode analysis.

Apart from the direct optimization algorithms mentioned above, solution scattering intensity guided molecular dynamics simulations are also used to sample structures that agree with the experimental data [48-50]. Molecular dynamic (MD) simulation requires forces on atoms. Thus the analytical derivative of the scattering intensity fitting potential is necessary in order to implement scattering information into an MD simulation. For Crysol, AXES and Fast-SAXS-pro introduced above, it is relatively difficult to obtain the analytical derivative of the scattering intensity with respect of the protein coordinates. This is because both the implicit and explicit hydration terms used in those methods are related to the protein coordinates through complicated functions. As an approximation, Bjorling et al. [48] and Kimanius et al. [49] consider only the scattering intensity from protein molecules, ignoring the contribution of the solvation layer, in their SAXS-guided MD simulations or meta-dynamics simulations. Their results are hardly comparable with the experimental data directly. Chen et al. [50] proposed a formula to compute the analytical derivative of the scattering intensity from all-atom explicit MD

simulations trajectories. The scattering intensity fitting potential is thus able to be added on the fly. The advantage of the method is the computation of the scattering profile at relatively large maximum q , up to $1.0 \text{ \AA}^{-1}$ due to the atomistic resolution of the MD simulation as well as the incorporation of explicit hydration water molecules. However, the computation is relatively time consuming based on exactly the same reason.

1.4 Outline of the research work

This work mainly focuses on improving the computational efficiency of current methods to evaluate the scattering intensities from protein structures, while keeping their accuracy. This is achieved by either improving the algorithm for the atomistic approach or fitting a set of more accurate coarse-grained form factors. Additionally, an implicit hydration term that is easily differentiable with respect to the protein coordinates is also developed. This can be used in evaluation of scattering intensities for large systems as well as structure optimization using derivative-based minimization algorithms. The three parts of research work in this thesis are briefly introduced as follows.

In the first part of the work (Chapter 3), a novel method to evaluate the scattering profiles of biomacromolecules based on the three-dimensional fast Fourier transform (FFT) is proposed. The protein solution system in SXS modeling typically contains a large number of atoms (or particles in the coarse-grained case), especially those including explicit hydration water molecules. Additionally, the scattering intensities will be evaluated for many times in the procedure of structure modeling. This makes the computational efficiency critical if the method is going to be used in the structure modeling

process. The FFT algorithm has already been implemented in the computation of SAXS and WAXS intensities in literature [51]. The Fourier transformation usually starts from three-dimensional electron density maps. In this part of work, we have developed a method to calculate the intensity directly based on the atomic coordinates of the system and the atomic form factors only. There is no need to convert the coordinates back to the electron density map in our approach. Because the atomic coordinates are widely available from Protein Data Bank (PDB) or MD simulations, this approach is particularly convenient for the computation of biomacromolecular systems. Furthermore, the approximations introduced when converting atomic coordinates to the electron density maps are avoided here. We implement our FFT-based method to improve the orientational average for the MD-based WAXSiS [33] method and the non-MD-based AXES [31] approach. Both methods consider the contribution of hydration water explicitly. The results are then compared with those from the non-FFT based online web servers. It is found that the accuracy of our method is comparable with the corresponding online servers, while the computation time becomes superior when the system contains more than approximately 3000 atoms. The biggest advantage of our method is the high computational efficiency and it should be noted that our FFT algorithm is generally applicable to other kinds of methods that employ atomistic coordinates in the scattering intensity computation.

In the second part (Chapter 4), we focus on improving the accuracy of coarse-grained form factors. CG form factors are usually parameterized by the IBA or IBA-like method in literature, as discussed in the previous parts. However, the accuracies of such methods are not satisfying for highly coarse-grained models. The electron density matching (EDM) optimization

scheme is proposed to systematically parameterize the coarse-grained amino acid form factors from their atomistic representations. A non-redundant protein structure dataset is constructed to perform the EDM optimization. The most essential part of our EDM method is the optimization through the least square fitting of the Fourier transforms of the protein electron densities. The reason to choose the fitting target instead of the scattering intensities is that the Fourier transforms can be expressed as the linear combination of form factors. Thus, by applying the least square fitting procedure, a mathematically guaranteed optimal solution can be obtained. This is advantageous compared with the non-linear optimization scheme used by Stovgaard et al. [42]. Besides, the resulting form factors can be easily separated to two components counting for the scattering of protein atom in vacuum and the scattering of protein excluded volume respectively. This allows us to optimize them separately and the contribution of the protein excluded volume effect can be tuned during the fitting to experimental scattering profiles. We tested our EDM method on both one-bead and two-bead per-residue CG schemes and compared our results with that of the IBA method. It is shown that CG SAXS profiles computed from our method is much better than those from the IBA method. Moreover, it is illustrated that the CG form factors derived from our EDM method can be used to compute theoretical SAXS profiles together with an explicit-solvent hydration layer around the protein. The obtained theoretical SAXS profiles can fit the corresponding experimental data very well.

In the third part of our work (Chapter 5), we aim to find a way to compute the overall protein scattering intensity, whose analytical derivative with respect to the protein coordinates can be easily obtained. This is because the availability of analytical derivatives is very important in the structure modeling based on

experimental scattering intensities, for both direct structure optimization method and SAXS-biased MD simulations. A simple implicit hydration term is introduced to modify the form factors of protein atoms or CG beads to reproduce the contribution of the protein hydration layer, as inspired by FoXS developed by Schneidman-Duhovny et al. [21, 22]. The implicit hydration term is proportional to the solvent accessible surface area (SASA) of the corresponding protein atom or CG beads. And the constant pre-factor is obtained by the optimal fitting of the overall scattering intensities computed from atomistic models together with explicit hydration water molecules. As the SASA can also be analytically evaluated from the protein coordinates, the scattering intensity is thus only related to the coordinates of protein atoms or CG beads, with its analytical derivative readily available. It is shown that the obtained implicit hydration scattering function can accurately reproduce the overall SAXS profiles of atomistic models. Furthermore, the coefficients in the implicit hydration function is shown to be similar for different levels of coarse-grained schemes, which means that our approximation, that the contribution of the hydration shell should be mainly related to the SASA of protein atoms, is valid at the SAXS region. We then performed structure optimization using the conjugate gradient algorithm. Final structures from optimization are close to our optimization target for four benchmark protein molecules. This indicates that the implicit hydration term developed in this work is of great potential to be used in future structure modeling based on experimental SAXS profiles.

2 General methodologies

2.1 Atomic form factors and the Debye formula

The X-ray scattering intensity of an object is computed from the square of the Fourier transform of its electron density distribution. In biological studies, the targets of scattering experiments such as protein molecules are usually represented by a set of discrete atoms for convenience. Thus it is favorable to use atomic form factors, which are the scattering amplitudes of isolated atoms, together with the atomic coordinates to compute the scattering intensity instead of converting them back to the electron density distribution. The atomic form factor of a certain atom is evaluated from the Fourier transform of its spherical electron density distribution, which can be computed from quantum chemistry by solving the Schrödinger equation. In the practical use, the atomic form factors are usually well approximated using the summation of four Gaussian functions as shown in equation (2.1).

$$f(q) = \sum_{k=1}^4 a_k \exp\left[-b_k (q / 4\pi)^2\right] + c \quad (2.1)$$

where the Cromer-Mann parameters a_k , b_k and c can be obtained from the *International Tables for Crystallography* [52].

As mentioned previously, the excluded volume effect is usually treated by the Gaussian sphere approximation [36] in literature. This can be fulfilled conveniently by modifying the atomic form factor adding an excluded volume correction as shown in the following equation.

$$f^{excl}(q) = f(q) - v\rho_0 \exp(-\pi v^{2/3} q^2) \quad (2.2)$$

where v is the van der Waals volume of the protein atom and ρ_0 is the electron density of bulk solvent, usually chosen to be 334 e/nm^3 . The atomic form factors of several common atoms, as well as their corresponding counterparts with the excluded volume correction are shown in Figure 2.1.

Figure 2.1 Atomic form factors for typical atoms that exist in protein molecules. The solid lines represent the original form factors, as computed in equation (2.1). And the dashed lines show corresponding form factors with the excluded volume correction, shown in equation (2.2).

Having the atomic form factors, the scattering amplitude of the protein molecules in vacuum can be evaluated by the Fourier transform shown in equation (2.3).

$$A(\mathbf{q}) = \sum_j f_j(q) \exp(-i\mathbf{q}\mathbf{r}_j) \quad (2.3)$$

And the scattering intensity can be calculated using the Debye formula [16] shown in equation (2.4).

$$I(q) = \sum_i \sum_j f_i(q) f_j(q) \frac{\sin(qr_{ij})}{qr_{ij}} \quad (2.4)$$

where i and j loops over all the atoms in the protein molecule and r_{ij} is the distance between the i^{th} and j^{th} atom. The Debye formula computes the theoretical orientational average based on the condition that the protein molecule is freely rotated in the solution state. The Debye formula is computationally expensive, as the formula is a double summation of the protein atoms with a computational complexity of $O(N^2)$. Moreover, as the Debye formula calculates the intensity instead of the amplitude, it is hard to handle the implicit treatment of the protein hydration layer based on the additivity of scattering component amplitudes, which is used in many existing programs in literature, such as Crysol. However, there still exist some methods that use the Debye formula to evaluate scattering profiles. They usually adopt a coarse-grained representation of protein structures and explicit hydration water molecules to reduce the number of atoms in the system, such as Fast-SAXS-pro. Nevertheless, the computational efficiency can be improved by advanced methods to perform the orientational average that will be discussed later.

2.2 Spherical average method to compute the orientational average

As shown previously, the computation of the one-dimensional scattering intensity always needs the orientational average over the scattering vector $\mathbf{q}$, for example, in equation (2.7) and (2.8). This is because the solute molecules can

rotate freely in the solution state during the scattering experiment. The rotation of solution molecule coordinates $\mathbf{r}$ is equivalent to the rotation of the scattering vector $\mathbf{q}$. Thus, in real computations, orientational average is performed over $\mathbf{q}$. The Debye formula shown in equation (2.4) is the theoretical average over all possible $\mathbf{q}$ directions given the scalar value of $\mathbf{q}$. However, the Debye formula is computationally expensive for relatively large systems. Thus, a more advanced method to perform the orientational average is proposed. In the work of Park et al. [32], Chen et al. [50] and Grishaev et al. [31], the spherical average method is used. In this method, the numerical average of the intensities at different orientations of $\mathbf{q}$ is performed instead of the analytical average. For each scalar value of q , a set of J directions are uniformly selected on the sphere with the radius of q in the reciprocal space. Different schemes can be used to select the orientations of $\mathbf{q}$. For example, the spiral method is used by Park et al. and Chen et al., while the Fibonacci grid method is used by Grishaev et al.. From our own tests, both methods are very accurate when the number of directions are as large as 1000, as used in their works. For the spiral method, the spherical coordinates of the J points on the q sphere are shown in the following equations.

$$\theta_j = \arccos \frac{2j-1-J}{J}, \quad \varphi_j = \sqrt{\pi J} \frac{2j-1-J}{J} \quad (2.5)$$

And the spherical coordinates of the Fibonacci grid method can be evaluated based on the following approach.

$$\theta_j = \arccos \frac{2j-1-J}{J}, \quad \varphi_j = (3-\sqrt{5})\pi j \quad (2.6)$$

where j loops over 1 to J in the above equations.

2.3 Evaluation of the scattering intensity based on single structure

In many approaches mentioned in the introduction part, the scattering intensity is evaluated from the input structure using equation (2.7) as follows.

$$I(q) = \left\langle \left| A_p(\mathbf{q}) - A_e(\mathbf{q}) + A_s(\mathbf{q}) \right|^2 \right\rangle_{\Omega} \quad (2.7)$$

In the above equation, $A_p(\mathbf{q})$ is the scattering amplitude of the protein in vacuum that is readily calculated using equation (2.3) shown in the previous part. $A_e(\mathbf{q})$ is the scattering amplitude of the excluded volume of the protein. $A_s(\mathbf{q})$ is the scattering amplitude contribution from the excess hydration layer and it is usually computed by scaling the scattering amplitude of the hydration layer with a small relative weight of those hydration water molecules. The latter two terms can be evaluated using both implicit and explicit solvent approaches as seen in literature. For example, in order to compute the scattering of protein excluded volume $A_e(\mathbf{q})$, the Gaussian sphere approximation, which is an implicit approach, is used by Crysol and many other methods. In this approximation, the excluded volume of each protein atom is supposed to be filled with electron gas in order to compute the term $A_e(\mathbf{q})$. The total number of electrons is the electron density of bulk solvent (usually chosen to be $0.334 \text{ e/\AA}^3$ following the experiment value) multiplies the van der Waals volume of the protein atom and the density of electron gas follows Gaussian distribution. For convenience, the modified atomic form factors computed by equation (2.2) commonly consider the excluded volume effect implicitly, where the first two terms can be computed together. The implicit treatment of the hydration layer varies between different methods. For example, Crysol uses a uniform electron

density of the hydration layer with the thickness of 3 Å to mimic the contribution of excess hydration waters, while FoXS simply incorporates the form factors of hydration water molecules into protein surface atoms. In both methods, the electron density or the number of hydration water molecules are tuned in order to obtain the best fitting to target SAXS profiles. Alternatively, the displaced solvent term and the hydration layer term can be computed using explicit water molecules whose positions can be determined by overlapping the protein structure with a pre-equilibrated water box, as done by methods like AXES.

2.4 Evaluation of the scattering intensity based on MD trajectories

Another type of methods computes the scattering profiles using two simulation trajectories, the protein solution simulation and the pure solvent simulation. This is to mimic the solution X-ray experiments, where the scattering intensity is measured for both the protein solution and the buffer system. The final scattering intensity is obtained by subtracting the scattering of buffer from the intensity of the protein solution.

In computational modeling, it is impossible to compute the scattering intensity of infinite systems directly. Thus envelopes are constructed around the protein structures for both trajectories. With some small approximations, it is shown that the scattering intensity is evaluated through the Fourier transform of the electron densities inside the envelopes [32-35]. Constructing a spherical envelope is the most straightforward way. This is used by previous works by Oroguchi et al. [34] and Kofinger et al. [35]. In other works by Park et al. [32] and Chen et al. [33], the envelope of the protein solution system is constructed

by including the protein molecule and all the water molecules within a certain distance d_{shell} to the surface of the protein. Similarly, a protein image is placed in the pure solvent system and the envelope is constructed by including the water molecules that is within the distance d_{shell} to the protein image. This kind of envelope reduces the number of atoms inside and is able to improve the computational efficiency.

Let A and B represent the envelopes of the protein solution system and the pure solvent system. $A(\mathbf{q})$ and $B(\mathbf{q})$ represent the Fourier transforms of the electron densities inside the corresponding envelopes, which can be computed by equation (2.3). Following the derivation of earlier works [32], the scattering intensity can be computed by the following equations.

$$I(q) = \langle D(\mathbf{q}) \rangle_{\Omega} \quad (2.8)$$

and

$$D(\mathbf{q}) = \langle |A(\mathbf{q})|^2 \rangle_{\omega} - \langle |B(\mathbf{q})|^2 \rangle_{\omega} - 2\text{Re} \left[\langle B^*(\mathbf{q}) \rangle_{\omega} \langle A(\mathbf{q}) - B(\mathbf{q}) \rangle_{\omega} \right] \quad (2.9)$$

In the above equations, $\langle \dots \rangle_{\Omega}$ represents the orientational average over different scattering vector $\mathbf{q}$ with the same scalar value due to the free rotation of solute molecules, while $\langle \dots \rangle_{\omega}$ denotes the ensemble average over the solute conformational change and solvent degrees of freedom. The third term in equation (2.9) means the real part of the complex number computed from the expression and “*” represents the conjugate complex number of the original. It is worth noting that in equation (2.9), the difference of the scattering intensities of the two infinite systems can be computed by only using the scattering

amplitudes inside the two envelopes through some reasonable approximations. The derivation of equation (2.9) is referred from the work by Park et al. [32].

2.5 Comparison between scattering profiles

In order to measure the discrepancy between the theoretically calculated scattering profiles (I_{cal}) and the experimental data (I_{exp}), the quantity chi-square (χ^2) defined in the following formula is minimized.

$$\chi^2 = N_q^{-1} \sum_{i=1}^{N_q} \left[\frac{I_{exp}(q_i) - fI_{cal}(q_i) + c}{\sigma(q_i)} \right]^2 \quad (2.10)$$

In the above equation, N_q is the number of points in the scattering intensity curve. f is the parameter to scale the theoretically evaluated scattering profile in order to fit the experimental one and c is the parameter that can reduce the error from the experimental buffer subtraction process. $\sigma(q)$ is the standard error of the experimental data at value q . Equation (2.10) is the general expression of χ^2 values and smaller χ^2 values represent the better overlapping of the comparing curves. In following chapters, we use slightly different expression of χ^2 values in different cases. When we compare the theoretical curve evaluated from MD simulation trajectories with the experiment, both parameter f and c are used to tune the fitting between experimental and calculated curves. This is because the computation from MD simulation trajectories is determined by the force field, without any freely tunable parameters, thus the experimental error of the buffer subtraction process must be considered. When the theoretical scattering intensities are computed from an isolated protein structure using equation (2.7), only the parameter f is used to minimize the χ^2 values. This is because during

this type of computation, the protein excluded volume as well as the weight of the hydration shell can be tuned to obtain the best fitting to experimental curves. Thus the error caused by the buffer subtraction should already be considered by tuning these two free parameters during the SAXS profile computation. The computed and experimental scattering intensities are usually presented at different q intervals, where experimental ones are usually in higher resolution. Thus, before comparing the two sets of scattering intensities, the computed scattering intensity is interpolated to the same resolution as the experimental data using cubic splines on the scale of $\log I_{\text{cal}}(q)$.

Additionally, when checking the accuracy of the fitted CG form factors and the implicit hydration term obtained in our works, the theoretical scattering intensities computed from atomistic protein structures are used as the comparison target. We still use a formula similar to the above to measure the discrepancy between the CG and atomistic SAXS profiles. However, as both curves are from theoretical computations, there is no need to perform the scaling by the parameter f , we simply set $f=1.0$. And the parameter c is also not used as the curves are not related to an experiment at all. At last, there is no standard error from the theoretical evaluated scattering intensities, thus we use an empirical standard error following Stovgaard et al. [42]. The resulting values are usually smaller compared with the χ^2 values between theoretical and experimental scattering profiles. In order to distinguish this with the χ^2 value define above, we use a new symbol R defined by the following equation.

$$R = N_q^{-1} \sum_{i=1}^{N_q} \left[\frac{I_{AA}(q_i) - I_{CG}(q_i)}{\sigma(q_i)} \right]^2 \quad (2.11)$$

Here the empirical standard error of the theoretically computed scattering intensities are approximated by

$$\sigma(q_i) = I(q_i)(q_i + \alpha)\beta \quad (2.12)$$

where α is chosen to be 0.15 and β is 0.3 following the work of Stovgaard et al. [42]. The reason that the computed R values are usually much smaller than χ^2 values may be the relatively large fluctuations observed in the experimental data.

2.6 Construction of non-redundant protein structure dataset

In order to parameterize CG form factors as well as the implicit hydration term, a non-redundant dataset containing high-resolution protein structures is extracted from the protein data bank (PDB) using the PISCES server [53] in our work. Only those crystal structures with resolution higher than 1.8 Å, R factor smaller than 0.15 and no missing residues are chosen. The PDB structures are culled by chains, and the sequence identity between different chains is filtered to be smaller than 10% in order to remove the redundancy. This results in 1003 single domain structures in our dataset (as of June 2015). For post-processing, the missing side chain residues are completed using the function module “complete_pdb” in Modeller [54]. Structure clashes are further removed by performing a short energy minimization using Gromacs 4.5.5 [55] and the CHARMM force field [56]. When parameterizing the implicit hydration term in Chapter 5, the protein structures in the dataset are further refined by removing those proteins with irregular shapes. The protein structures with radius of gyration (R_g) larger than 1.15 fold of their standard R_g are then removed from

our dataset. The standard R_g of a protein with a globular shape is estimated by the empirical formula $R_g = 0.395N^{3/5} + 7.257$ [57]. The remaining database to perform optimization of the implicit hydration term contains 880 single domain protein structures.

2.7 Protein coarse-grained mapping schemes

In Chapter 4 and 5, the coarse-grained protein structures are used in the parameterization of CG form factors as well as the implicit hydration term and the computation of CG SAXS profiles. Two categories of CG mapping schemes, one-bead per-residue and two-bead per-residue, are adopted. Apart from the commonly used coarse-grained mapping to place the CG beads at the position of $C\alpha$ and center of mass (COM), we also test the new center of electrons (COE) coarse-grained scheme, in which the position of the CG beads are evaluated through the following equation.

$$\mathbf{R}_{CG} = \frac{\sum_j e_j \cdot \mathbf{r}_j}{\sum_j e_j} \quad (2.13)$$

Here, e_j is the number of electrons contained in atom j , while r_j is the Cartesian coordinate of atom j . j loops over all the atoms inside the CG group. This COE scheme is proposed on the basis of the physical meaning of the CG form factors, which is the Fourier transforms of the electron densities of the corresponding coarse-grained beads. So fitting the CG form factors is actually equivalent to the fitting of CG electron densities to atomistic electron densities. Because the CG form factors are one-dimensional functions of the scalar q value, the

electron density of the CG bead is supposed to be spherically symmetric in the real space. Thus, placing the CG bead at the center of electrons can ensure relatively better overlapping between the electron density distributions of the atomistic and CG representations. In summary, we have in total four types of CG schemes in two coarse-grained levels. For the one-bead per-residue CG scheme, the CG bead of each residue can be placed at its $C\alpha$ position or the COE of the corresponding residue. For the two-bead per-residue CG model, one CG site is used to represent the backbone of the residue (containing N, $C\alpha$, C, O, HN and $H\alpha$ six atoms) and another representing the side-chain (containing all the rest atoms). When the COE mapping scheme is used, the two CG beads are placed at their corresponding center of electrons. When the $C\alpha$ and COM mapping scheme is used, the CG bead representing the backbone is placed at the $C\alpha$ position of the residue, while the CG bead representing the side-chain is placed at the COM of the side-chain atoms.

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

3.1 FFT algorithm implemented to orientational average

Theoretical orientational average computed by the Debye formula is computational expensive, as discussed in previous part. Although the improved numerical spherical average method has a better computational complexity of $O(N)$, it can still take significant calculation times, especially for methods using atomistic representation of protein structures combined with explicit hydration solvent treatment. In order to improve the computational efficiency of the scattering intensity, the fast Fourier transform (FFT) algorithm is implemented to compute the scattering amplitude $A(\mathbf{q})$ in our work. Compared with the spherical average method, the advantage of the FFT algorithm is that the $A(\mathbf{q})$ values at all the grid $\mathbf{q}$ points can be evaluated at once from discrete electron densities $A(\mathbf{r})$.

From equation (2.3), the FFT algorithm cannot be directly applied to compute $A(\mathbf{q})$, because the right side of the equation contains additional term $f_j(q)$, which is related to both the atom type and the scalar value of $\mathbf{q}$. So we first split our system into several subsystems, each containing individual type of atom. $A'(\mathbf{q})$ is then evaluated for each subsystems, which is defined by the following equation.

$$A'(\mathbf{q}) = \sum_j^{N'} \exp(-i\mathbf{q}\mathbf{r}_j) \quad (3.1)$$

where the superscript t means the type of atom in this subsystem and N^t is the number of atoms belongs to this type. Combining equation (2.3) and (3.1), $A(\mathbf{q})$ can be expressed as the summation of the form factor $f^t(q)$ multiplies $A^t(\mathbf{q})$, as shown follows.

$$A(\mathbf{q}) = \sum_t A^t(\mathbf{q}) f^t(q) \quad (3.2)$$

$A^t(\mathbf{q})$ can thus be discretized to $\mathbf{q}$ grid points in the reciprocal space, and computed by the FFT algorithm at once. We construct cubic grids in both the $\mathbf{q}$ space and $\mathbf{r}$ space. In order to rewrite the expression of $A^t(\mathbf{q})$ into the format of discrete Fourier transform, the bin width of the cubic grid and the maximum value of q and r must satisfy the following relationship.

$$r_{bin} = \frac{2\pi}{2q_{max}}, \quad q_{bin} = \frac{2\pi}{r_{max}} \quad (3.3)$$

The reason that $2q_{max}$ is used in the first part of the formula is because the average of $\mathbf{q}$ to scalar value q includes negative $\mathbf{q}$ points. The range of reciprocal space on each dimension should be $-q$ to q . We compute the particle density at each $\mathbf{r}$ grid point by discretize the atomic coordinate $\mathbf{r}_j$ to its nearest neighboring grid point. The equation (3.1) can thus be rewritten in the format of discrete Fourier transform as the following.

$$A^t(h, k, l) = \sum_a \sum_b \sum_c \left\{ \rho(a, b, c) \exp(-2\pi i (h, k, l) \cdot (a, b, c)) / N_g^3 \right\} \quad (3.4)$$

In the above equation, the vector (a, b, c) represents the indices of grid points in the real space, while the vector (h, k, l) denotes the indices of the grid points in the reciprocal space. The three dimensions have the same number of grid points

N_g as we are using cubic grids here. $\rho(a, b, c)$ is the number density of atoms with type t at the corresponding grid point. The FFT algorithm can then be applied to the discrete Fourier transform problem shown in equation (3.4) to obtain all the $A'(\mathbf{q})$ values at once.

A major problem encountered here is that the obtained scattering amplitude $A'(\mathbf{q})$ and $A(\mathbf{q})$ are at very low resolution. This is because the size of the system is usually not very large. For example, a typical MD simulation box may be at the length scale of several nanometers for moderate size protein molecule simulations. According to equation (3.3), the obtained q_{bin} value will be relatively large. As an example, a simulation box of 10 nm will result in a q_{bin} value of approximately 0.6 nm^{-1} . The scattering intensity finally obtained with such a large interval is hardly comparable with any real experimental data, making the result useless. The zero padding method [58], which is widely used in the field of digital signal processing, is thus applied in order to solve the low resolution problem. Basically, the obtained $A(\mathbf{q})$ with relatively low resolution is first transformed back to the electron density $A(\mathbf{r})$. Then, we expand the original system to a much larger simulation box by filling zero electron density to the newly expanded region, as there are actually no atoms inside. The size of the new enlarged system is chosen based on the desired q_{bin} value. In this study, $q_{\text{bin}} = 0.2 \text{ nm}^{-1}$ is selected, which corresponds to the new r_{max} value of about 31.4 nm. The scattering intensity at this resolution is enough to represent the necessary features and further increase to the resolution will need additional computational time. We denote the electron density of the enlarged system to be $A(\mathbf{R})$, and finally the Fourier transform of $A(\mathbf{R})$ is computed to obtain the scattering amplitude $A(\mathbf{q})$ at a higher resolution. This Fourier transform step is the most time-consuming part in the whole algorithm. This is because the

number of grid points is relatively large due to the two reasons shown below. The first reason is the q_{bin} of the output must be small enough so that the final result is meaningful and comparable with the experimental scattering profiles as discussed previously. In the other aspect, we also need to make r_{bin} relatively small in order to ensure the error caused by assigning atoms to their nearby grid points in the real space is acceptable. From our testing, r_{bin} should be as small as about 0.1 nm in order to make the final calculation result accurate enough. In our computation, we select q_{max} to be 30 nm^{-1} or even larger according to the value of r_{bin} . This is far exceeding the value we actually needed as the typical q_{max} value is 15 or 20 nm^{-1} in standard WAXS experiments. It is worth noting that the increase of atom numbers and atom types will not affect the computational efficiency significantly, because the evaluation of number density distribution and atom-type based Fourier transform occupy only very small proportion of the overall computational time. It is shown that the computation efficiency of our FFT-based method is still much higher than the spherical average method widely used in literature for most computations with explicit solvent treatment.

As shown in equation (2.7) and (2.8), we always need to project the three-dimensional scattering intensity computed from the scattering amplitude $A(\mathbf{q})$ to the one-dimensional scattering intensity. We use a simple method called nearest neighbor searching to accomplish this process. In this method, the scattering intensity at each $\mathbf{q}$ grid point is added up to output scalar q value that is the nearest to its modulus. The scattering intensity value at each output q value is then normalized by the number of nearest $\mathbf{q}$ grid points. There are also methods such as channel sharing to perform the projection from three-dimension to one-dimension in literature. However, we find there is no

obvious improvement of the final result and the nearest neighbor searching method is still used in our work.

In the methods to evaluate scattering intensity from MD simulation trajectories, the MD simulation itself will also take quite a lot of computational time besides the calculation of scattering intensity. This makes the superiority of our FFT-based method not so significant over other methods. However, for the methods that does not need the expensive MD simulations, the computation efficiency will be benefited greatly from the FFT algorithm. Moreover, our FFT-based method can also be implemented to SAXS/WAXS guided MD simulations, such as the work by Chen et al. [50]. In these methods, the scattering intensity needs to be evaluated on the fly during the MD simulation, thus the improvement of computational efficiency for scattering profile evaluation will greatly reduce the overall computational time.

3.2 Details of MD simulations and our FFT-based method

Three proteins whose WAXS profiles are freely available in the Small Angle Scattering Biological Data Bank (SASBDB) [15] are chosen to be our benchmark systems. They are Lysozyme (PDB ID: 1LYZ) [59], Myoglobin (PDB ID: 1WLA) [60] and RNase (PDB ID: 1C0B) [61]. The crystal structures obtained from the Protein Data Bank (PDB, <http://www.rcsb.org>) are used as initial structures for MD simulations. We also completed the missing side chains of the crystal structures using the corresponding tool on the WHAT IF server [62]. The atomistic simulations with explicit solvent were performed using Gromacs 4.5.5 MD simulation package [55] and the CHARMM 27 force field [56]. We used very large dodecahedral simulation boxes in order to study

the envelope size effect. The nearest distance of the protein molecule to the edge of the box was set to be 5 nm, far larger than normal simulations. The resulting box size was about 15 nm. The TIP3P water model [63] was used to solvate the protein molecules. Finally, the system is neutralized using sodium or chlorine ions. Each system was equilibrated using a short constant NPT simulation of 100 ps. Then the constant NPT production run of 1 ns was performed. The simulation frames were collected every 10 ps during the production run, resulting in total 101 frames for all the simulation systems. Apart from the protein solution simulation, we also performed the pure solvent simulation with the exactly same box size in order to compute $B(\mathbf{q})$ needed in equation (2.9). Additionally, short protein solution MD simulations of 200 ps were performed with the protein backbone restrained, in order to study the effect of protein conformational change to the final computed scattering intensity. The short simulations were also performed at constant NPT condition.

After the MD simulations, two envelopes were constructed. One containing the protein molecule as well as nearby solvent water molecules to compute $A(\mathbf{q})$ and the other with pure solvent to compute $B(\mathbf{q})$. Because the computation efficiency of our FFT method remains nearly unchanged as the number of atoms in the system increases, we simply adopted a spherical envelope with the center of protein taken as the center of the sphere for computational convenience. The radius of the envelope is chosen to be $D_{\max}/2 + d_{\text{shell}}$, where $D_{\max}$ is the maximum diameter of the protein molecule and d_{shell} is the thickness of the solvent molecules included in the computation. However, when we compare the computation time with the spherical average method in literature, we use the same type of envelope as reported in order to make a relatively fair

comparison. In this way, only the solvent molecules whose nearest distances to the protein surface are smaller than d_{shell} are included in the envelope.

We also need to tune the electron density of the system in order to obtain more accurate scattering intensity curves. This is because the electron density of bulk water is usually slightly different from the value measured in experiment (334 e/nm^{-3}) due to the inaccuracy of force field parameters. The electron density of the water molecules inside the envelope is scaled by the ratio of $\delta_{A,B} = \rho_{\text{exp}} / \rho_{A,B}^{\text{bulk}}$ for both the pure solvent and the protein solution system. The scaling ensures the electron density of bulk water outside the envelope (ρ^{bulk}) in the simulation is tuned to the experiment observed value (ρ^{exp}).

3.3 Accuracy of the FFT-based method

As the first step, we performed fine-tune of the parameters in our FFT-based method in order to obtain the optimal computational efficiency as well as to maintain the reasonable accuracy. We chose the result computed from the Debye formula method as the standard curve for the comparison, as the Debye formula is the theoretical orientational average without any kind of approximations. We have compared the results of our FFT-based method with the spherical average approximation method presented in literature to the result of the Debye formula. The three benchmark protein molecules Lysozyme, Myoglobin and RNase without solvent were used to compute the scattering intensity curve in order to reduce the computational time. The R value defined in equation (2.11) is used to measure the difference between results from the FFT-based method or spherical average method and the standard profiles

evaluated from the Debye formula. Note that both scattering profiles used for comparison are from atomistic model instead of CG model, slightly different from the equation of the R value. The empirical standard error value of $\sigma(q_i) = I(q_i)(q_i + \alpha)\beta$ is used for the theoretical curves computed from the Debye formula method. Here $\alpha=0.15$ and $\beta=0.3$.

The variation of the R values between the scattering intensities computed from our FFT-based method and the Debye formula method for different $q_{\max}$ values is shown in Figure 3.1(a). From the figure, The R value decreases as the $q_{\max}$ value increase for all the three proteins. This is quite straightforward as suggested by the relationship $r_{\text{bin}} = 2\pi/(2q_{\max})$ discussed above, which means larger $q_{\max}$ values correspond to smaller r_{bin} values. When smaller r_{bin} values are used, the error of assigning the atomic coordinates to nearby $\mathbf{r}$ grid points will be reduced. On the other hand, the q_{bin} value is fixed to be 0.2nm^{-1} in our computation, thus using larger $q_{\max}$ values means more overall grid points in the Fourier transform process, which will result in higher computational cost. It is observed that the R values in Figure 3.1(a) becomes very small (4.496×10^{-2} , 5.235×10^{-2} and 4.686×10^{-2} for the three proteins respectively) when $q_{\max}$ is chosen to be 30 nm^{-1} . No obvious decreasing of R value is observed with the further increment of the $q_{\max}$ value. Thus in all the following computations, the value of $q_{\max}$ is selected to be 30 nm^{-1} , used together with the q_{bin} value of 0.2 nm^{-1} . Besides, the scattering intensity profiles computed using our FFT-based method with the selected parameters, the spherical average method and the Debye formula for the three benchmark proteins are shown in Figure 3.1(b). The scattering intensities from spherical average method overlaps with the Debye formula perfectly with extremely low R values (3.068×10^{-5} , 3.433×10^{-5} and 5.528×10^{-5} for the three proteins). Although the R values of our FFT-based

method is slightly larger than the spherical average method, the differences between scattering curves are almost indistinguishable within the line width.

Figure 3.1 (a) the R values between the scattering intensities computed from the Debye formula and our FFT-based method when different $q_{\max}$ values are used. (b) Overlapping of the scattering profiles computed from the Debye formula, the spherical average method and our FFT-based approach.

3.4 The computational efficiency of our FFT-based method

The computational times consumed by our FFT-based method are evaluated for systems containing different number of atoms and compared with the spherical average method. We used the 101 simulation frames obtained in the production simulation of all the three proteins as well as the pure solvent to evaluate the computational efficiency. The protein shaped envelopes are constructed following the spherical average methods in literature in order to have a fair comparison with same number of atoms in the systems. To include different number of atoms inside the envelopes, we use different values of d_{shell} , which is the thickness of the hydration layer, varies from 0.5 to 3 nm. We implemented both methods using our homemade version of Gromacs analysis tool. And the computation times are evaluated on single Intel Xeon CPU (E7-8850 with 2.0 GHz frequency). The resulting times are plotted in Figure 3.2, with the three benchmark proteins colored in red (Lysozyme), green (Myoglobin) and blue (RNase) respectively. The computational times obtained from our FFT based method are represented in cubic points, while the results of the spherical average method are shown by spherical points. Different points on each curve represent the different shell thickness ranging from 0.5 nm to 3.0 nm. The numbers of atoms for each system with different shell thickness are labeled in the x axis. It should be noted that we use different shell thickness in order to generate different number of total atoms inside the envelope. This is because there are no reliable WAXS data for very large protein molecule in the SASBDB database. When considering the computational efficiency only, small proteins with very thick hydration layer is basically equivalent to larger proteins with relatively thinner hydration layer.

Figure 3.2 The comparison of computational times consumed by the spherical average method and our FFT-based method for the three benchmark proteins. The computational times are estimated using 101 frames of MD simulation trajectories. Data points on each curve correspond to different hydration shell thickness used to construct the envelope. They are 0.5, 0.7, 1.0, 1.5, 2.0, 2.5 and 3.0 nm respectively. The results of Lysozyme, Myoglobin and RNase are colored in red, green and blue respectively. The spherical points represent the computational times of the spherical average method, while the cube points represent those of our FFT-based method.

From Figure 3.2, it is very clear that the computational time of the spherical average method increases linearly with the number of atoms included in the envelope. However, for our FFT-based method, the time consumed remains nearly unchanged when the system size increases. This result is consistent with our previous discussions. For the spherical average method, the computation complexity of the algorithm is $O(JN_aN_q)$. J is the number of scattering vectors used for each scalar value of q . N_a is the number of atoms in the system and N_q is the number of points in the output scattering intensity curves. In our test, J is fixed to be 1000 and N_q is fixed to be 51, where q ranges from 0 to 10 nm^{-1} with

an interval of 0.2 nm^{-1} . Although the q_{max} value in our FFT-based method is chosen to be 30 nm^{-1} , we only compute the scattering intensity to 10 nm^{-1} for the spherical average method for fair comparison. Thus the computation time should be linearly related to the number of atoms N_a , as shown in the figure. For our FFT-based method, the computational complexity is $O[N_{\text{FFT}}^3 \log(N_{\text{FFT}}^3)]$, where N_{FFT} is the number of grid points in each dimension after zero padding for both the reciprocal and real space. Because we use the q_{bin} value of 0.2 nm^{-1} , which corresponding to r_{max} of about 31.4 nm, so the number of grid points N_{FFT} is fixed to be $2q_{\text{max}}/q_{\text{bin}}=300$ for all the systems whose size are smaller than 31.4 nm. Thus the computational time of our FFT-based method remains nearly unchanged for those systems. Increasing of number of atoms will slightly increase the time to evaluate the number density of atoms on each real space grid points. However, this time is negligible. From Figure 3.2, we can find the number of atoms when the computational efficiency of the spherical average algorithm is approximately same with our FFT-based algorithm. It is estimated to be 3000 atoms and the FFT-based method becomes superior when more atoms are included in the system. This is a relatively small number for atomistic protein structures combined with explicit solvent molecules. For example, the atomistic representation of protein Lysozyme contains 1960 atoms. When a protein shaped envelope with 0.5 nm thickness hydration layer is constructed, the total number of atoms contained in the envelope will be around 4444. This is already larger than the value estimated above. Thus, for most applications of WAXS profile computation, where the atomistic structure details of both protein and hydration water molecules are usually required to ensure the accuracy, our FFT-based method will be advantageous for the computational efficiency.

As discussed above, the size-independent computational efficiency of our FFT-based method applies for systems whose size is smaller than 31.4 nm, including the protein molecule and the hydration shell inside the envelope. To have such a large system, the protein molecule is estimated to have a molecular weight of approximately 10^4 kDa, supposing the protein is near globular with the average density to be 1.22g cm^{-3} and the hydration shell thickness is chosen to be 0.7 nm. This molecular weight is even larger than the whole complex of eukaryotic ribosome, which is about 3.2×10^3 kDa. So the system size is large enough for the MD simulations of commonly available proteins. When the size of the computed molecular system becomes larger than the value of 31.4 nm, no further zero padding is needed. Thus the number of FFT grid points along each dimension will increase following the system size. Then the total number of grid points N_{FFT}^3 becomes approximately proportional to the number of atoms N_a in the molecular system. The computational complexity of the FFT-based method are then reevaluated as $O[N_a \log(N_a)]$ approximately. It should be noticed that the pre-factor is much smaller than that of the spherical average method, from the computational time comparison shown in Figure 3.2.

Additionally, during the comparison of computational efficiency, we constructed envelopes based on the protein shape to reduce the total number of atoms inside the envelope, in order to have a fair comparison with the spherical average method. However, because our FFT-based method is insensitive to the number of atoms in the system up to a very huge system size, we can simply use a spherical shape envelope in our computation. This will further saves some computation times when determining which solvent molecules are inside the envelope, which can be relatively time consuming for relative large size protein molecules.

We also implemented the FFT-based algorithm to the non-simulation based method to calculate the scattering intensity where explicit hydration solvent molecules are also used. The method AXES [31] is improved with the FFT algorithm discussed above and the computational efficiency is compared with the original method. We still use Lysozyme as the benchmark system. The result of AXES is an average over the scattering intensities computed from protein molecule overlapped with all the 101 frames of equilibrated water box. The computational time is tested using our own implementation for both the original and FFT-based AXES method. The orientational average of the original AXES method is chosen to be the Fibonacci grid method with 1589 directions used for each scalar value of q , and the thickness of hydration layer is set to be 0.3 nm, following the work of Grishaev et al. [31] and the value of $\delta\rho = 0.04$ is used in our computation for best overlapping with the experimental scattering profile. The molecular system contains about 6500 atoms, including the protein molecules, displaced water molecules and hydration water molecules. Implementation of the FFT algorithm has significantly reduced the overall computation time from 1290 s to 337 s, an increment of about four fold.

3.5 Comparison of the scattering profiles with experimental data

In this section, we compared the scattering profiles evaluated using our FFT-based method with the experimental data obtained from SASBDB, for the three benchmark proteins (Lysozyme, Myoglobin and RNase). Experimental data points at small q region are not very accurate, as shown by the non-linear Guinier plots of the scattering intensities and are thus discarded in our comparison. Only the experimental scattering profiles start from 0.309, 0.248

and 0.631 nm^{-1} , respectively, are used here for the three proteins. The distance of the envelope to the protein surface was set to be 0.7 nm following the value used by Park et al. [32] and Chen et al. [33]. We also computed the scattering intensities of the three proteins using the WAXSiS server [33], where no FFT-algorithm is applied, using exactly the same parameters as a comparison. As shown in the method part, the computed scattering profiles are interpolated to have the same resolution as the experimental data before performing any comparison. The computed curves are then fitted to experimental ones by minimizing the χ^2 values defined in equation (2.10).

Table 3.1 The χ^2 values of the computed scattering profiles calculated by the Waxis server and our FFT-based method, compared with the experimental scattering intensities of the three benchmark proteins.

	Lysozyme	Myoglobin	RNase
Waxis	0.515	1.192	0.848
FFT	0.564	2.130	0.504

Figure 3.3 The scattering intensities computed using the atomistic MD simulation trajectories by the WAXSiS server and our FFT-based method, compared with the experimental scattering profiles after minimizing the χ^2 values for the three benchmark proteins.

The resulting χ^2 values are listed in Table 3.1 and the overlapping curves are plotted in Figure 3.3. From the figure, the computed scattering profiles fit the experimental data very well, except there are some discrepancies in the fitting of Myoglobin. In the table, the χ^2 values are smaller than 1.0 for two of the three proteins, suggesting our computed scattering profiles can perfectly fit the scattering experiment results. For the first protein Lysozyme, it is shown the scattering profiles from both computational methods agree very well with those from experiment. The χ^2 value of the result from WAXSiS server is slightly smaller than that of our FFT-based method (0.515 compared with 0.564), and the major discrepancy between the result of our method and the experiment

appears at the middle q region of $6 \sim 7 \text{ nm}^{-1}$. For the protein myoglobin, the fitting to experimental result is the worst among all the three benchmark cases. The result of the WAXSiS server has a reasonable overlapping with the experimental one until about $q \approx 8.5 \text{ nm}^{-1}$ and the χ^2 value is 1.192, which is still a relatively good fitting result. On the other hand, the χ^2 value of our FFT-based method is slightly larger with a value of 2.130. The computed curve deviates slightly from the experimental one in the q region of $2 \sim 3 \text{ nm}^{-1}$ as well as $5.5 \sim 7.5 \text{ nm}^{-1}$. However, the overall shape of computed curve from our FFT method is still within acceptable agreement with the experiment. The relatively large χ^2 values for the protein Myoglobin is probably partially due to the large cofactor HEME that persists in the middle of the protein molecule. It is known that there exist a relatively large number of delocalized π electrons as a result of the porphyrin macrocycle. The interactions between the HEME and nearby atoms may not be able to describe by the classical MD force field properly. Additionally, from Table 3.2, it is shown the resulting scattering intensity fits the experimental profile with a significant larger χ^2 value when the simulation trajectory with protein backbone restrained is used to perform the computation. This indicates that the crystal structure of Myoglobin used as initial conformation of MD simulation maybe quite different from the real solution state protein conformations. This should also be one of the reasons of the large χ^2 values. For the last protein RNase, it is found that our result is even slightly better than that obtained from the WAXSiS server, as shown by the χ^2 values of 0.504 versus 0.848. From the figure, the calculated scattering intensity from the WAXSiS server overlaps with the experimental one quite well until about $q \approx 8 \text{ nm}^{-1}$, after which the computed curve values are somewhat larger than the experimental target values. For our FFT-based method, the tail region has a

better fit compared with the WAXSiS server, while values in the q region of around 7 nm^{-1} are overestimated on a very small scale.

Table 3.2 The χ^2 values between the computed and experimental scattering intensities for the three benchmark proteins using different solvation shell thicknesses. The protein backbone restraint MD simulation trajectories are used to compute scattering intensities.

Thickness of solvation shell (nm)	Lysozyme	Myoglobin	RNase
0.5	0.560	5.605	0.647
0.7	0.616	6.078	0.735
1.0	0.990	3.804	0.882
1.5	1.087	4.107	0.977
2.0	1.248	5.260	0.775
2.5	1.032	7.172	0.677
3.0	2.256	5.972	1.287

As a summary, the computed scattering profiles from our FFT-based method are comparable with those obtained by the WAXSiS server, when the scattering intensities from experiment are used as target. The observed differences between the results of the two methods may be caused by the somewhat different protein conformations presented in the MD simulation with explicit solvent. From our own implemented spherical average method, the results obtained are very similar with that of the FFT-based method using the exactly

same MD trajectories. There are only some small deviations shown in the relatively high- q region. Additionally, from Table 3.2, the χ^2 values of all the calculated scattering profiles becomes larger, in case of the protein backbone restrained MD simulation trajectories are used, indicating the protein conformational fluctuations can improve the accuracy of the resulting scattering profiles. This is consistent with the work by Chen et al. [33], as well as the fact that the experimentally measured scattering data is the average over the solution-state protein conformational ensemble. The improvement of including conformational fluctuations varies for different kinds of proteins, relating to the similarity between the protein crystal structure and its conformational ensemble in solution state.

Figure 3.4 The scattering intensities of Lysozyme computed using the FFT implementation of the AXES approach and the original AXES server, compared with the experimental data after minimizing the χ^2 values.

For the non-simulation based method, the comparison of the computed scattering profiles from the AXES server as well as our implementation of the same method using the FFT algorithm with the experimental data is shown in Figure 3.4 for the protein Lysozyme. The χ^2 values of both calculated curves are quite small, with our FFT implementation ($\chi^2=0.611$) slightly better than the result obtained from the AXES server ($\chi^2=0.745$). We have used a different $\delta\rho$ value with that of the AXES server (0.04 versus 0.02), as that value results in the best fitting to the experimental scattering intensity. This difference of $\delta\rho$ values and the χ^2 values should be mainly due to different MD trajectories of equilibrated water box used by us and the server.

It is also find that the final resulting χ^2 values are heavily related to the minimization formula used. We have tested the effect of removing the parameter c in equation (2.10). All the χ^2 values are shown to be worse than the current results. While the results of our FFT-based method are still somewhat acceptable, the new χ^2 values of the computed curves from the WAXSiS server become even worse. We conclude that the parameter c is necessary to remove the errors introduced in experimental buffer subtraction process, when the theoretical scattering intensities are computed *ab initio* from MD simulation based methods.

3.6 The effect of different envelope size

During the derivation of equation (2.9), there are some approximations that the size of envelope must be large enough so that it contains all the hydration layer water molecules and part of the bulk water molecules. Thus it is important to know which value of d_{shell} should be enough for common protein simulation

trajectories. In their previous work, Chen et al. [33] studied the electron density distribution along the distance to the protein surface averaged from the trajectories, as well as the effect of shell thickness to the radius of gyration (R_g) of the solute molecules and the scattering intensity at $q = 0$ for relatively small d_{shell} values up to 1.2 nm. In our study, a sufficient large simulation box is used to perform the MD simulation and the high computational efficiency of our FFT-based method allows us to compute the scattering intensity using much larger d_{shell} values. Furthermore, the overall fitting to the experimental data is examined, given a more comprehensive view of whether the shell thickness is adequate.

Table 3.3 The χ^2 values between the computed and experimental scattering intensities for the three benchmark proteins using different solvation shell thicknesses. The MD simulation trajectories without protein backbone restraint are used to compute scattering intensities.

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

Figure 3.5 (a) The scattering intensities of Lysozyme computed using different solvation shell thickness ranging from 0.5 to 3.0 nm. (b) The comparison of the scattering intensities from different solvation shell thickness with experimental scattering intensities.

Six shell thickness values ranging from 0.5 to 3.0 nm were chosen to evaluate the scattering profiles using the MD simulation based method for all the three benchmark proteins. The results for protein Lysozyme is shown in Figure 3.5, with the direct comparison of scattering intensities plotted in Figure 3.5(a) and

the fitting to experimental data illustrated in Figure 3.5(b). From the figure, it can be seen the scattering profiles are very similar at region with small q values when different shell thickness values are used, except the curve of $d_{\text{shell}} = 3.0$ nm. At high q region, the scattering intensities become slightly different. There is no significant trend of convergence when the larger d_{shell} values are used in the computation. It should be noted that the resulting scattering intensity profile computed using 3.0 nm shell thickness is significantly different from other profiles for both low q and high q region. The observed inaccuracy in this case may be originated from the density correction process during the evaluation. Because a very huge envelope is used, the number of water molecules left outside the envelope may be insufficient to correctly evaluate the bulk water density. The water density fluctuations sampled in the MD simulation trajectories could affect our final calculation result both at $q = 0$ and the whole curves, when different envelope sizes are chosen. However, it is shown in Figure 3.5(b) that the fitting to experimental data is not obviously affected by the chosen of different shell thickness values except the largest one. The χ^2 values of computed scattering intensities using different envelope sizes for the three benchmark proteins are listed in Table 3.3 and it is shown that all the shell thickness values ranging from 0.5 to 2.5 nm can result in reasonable fitting to the experimental data. Generally, the scattering intensities calculated using shell thickness values of 1.0 or 1.5 nm have slightly better overlapping with experiment. Nevertheless, the small differences between different envelope sizes are negligible due to the existing errors in the experimental scattering profiles. In summary, our computation is consistent with those in previous studies, showing the recommended shell thickness value of 0.7 nm is reasonable.

3.7 Conclusions

Although there is a large number of SAXS/WAXS experimental data in structural biology, the one-dimensional scattering intensity curves contain only degenerated structural information, due to the freely translational and rotational movements of molecules in solution state. The computational tools that can evaluate the theoretical scattering intensity of a given structure are the first step toward structure modeling using the experimental scattering data. In our study, we implemented the FFT algorithm to both the MD simulation based and non-simulation based methods that can accurately compute the scattering profiles with improved efficiency compared with previous studies reported in literature. The method can contribute to extracting structural information from the experimental SAXS/WAXS profiles. When applied to the MD simulation based method, our method does not require additional parameters to compute the scattering intensities and the structure and density of the hydration layer as well as displaced solvent molecules are purely determined by the molecular force field.

While the computational efficiency of our FFT-based method is already much higher than the spherical average method for moderate size protein solution simulation systems, the computational time will remain nearly unchanged for increasing system size until $r_{\max}$ reaches about 31.4 nm. This will greatly benefit the computation for very large systems, such as multiple protein complexes as well as molecular processes like protein aggregation. Other possible application of our method may be the MD-based and SAXS-guided structure optimization.

The source codes and executable programs of all the method used in this part of work are freely available at <http://www.ntu.edu.sg/home/lylu>, including our implementations of both the spherical average algorithm and the FFT algorithm for MD-based as well as non-simulation based methods. Simple examples with sample input and output files can also be accessed at the above address.

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

4.1 Methods to derive amino acid form factors

Coarse-grained representations are desired in evaluating SAXS profiles from protein structures. This is because the nature of SAXS is mainly determined by the low resolution structure shape. Practically, this CG representation can greatly improve the computational efficiency by reducing the number of particles. Toward this direction, there are already many methods, such as isolated bead approximation (IBA) and IBA-like methods, to parameterize coarse-grained form factors. However, they generally result in large errors, especially at residual coarse-grained level. Thus, in this study, we have proposed an optimization scheme, termed the electron density matching (EDM) method, which can be used to derive the coarse-grained (CG) form factors of amino acid residues from their atomistic structure coordinates. The EDM method is basically inspired by the force-matching method, which is used to optimize the CG force field parameters in molecular modeling.

4.1.1 Electron density matching (EDM) method

In our EDM method, the optimization target is selected to be the scattering amplitude, which is defined as the three-dimensional Fourier transforms of protein electron density. The scattering amplitude can be evaluated from both the CG and atomistic representation of a given protein structure, and the difference between the results from two different resolutions are minimized to

obtain the CG form factors. Our optimization method is different from the previous work by Stovgaard et al. [42], in which the one-dimensional scattering intensity $I(q)$ is chosen to be the fitting target. The Fourier transform of protein electron density contains additional information prior to the orientational averaging process to compute $I(q)$, thus our method should be more rigorous. Because the three-dimensional Fourier transform of the real space is unitary, the electron density distribution of the target protein should be faithfully reproduced if its Fourier transforms over all the possible scattering vectors $\mathbf{q}$ are well fitted.

Specifically, the Fourier transformation of protein electron density at a certain scattering vector $\mathbf{q}$ can be evaluated from both the atomistic and coarse-grained representation of the protein structure using the following two formulas respectively.

$$A(\mathbf{q}) = \sum_j f_j^{excl}(q) \exp(-i\mathbf{q}\mathbf{r}_j) \quad (4.1)$$

$$A(\mathbf{q}) = \sum_J F_J^{CG}(q) \exp(-i\mathbf{q}\mathbf{R}_J^{CG}) \quad (4.2)$$

Equation (4.1) is repeated here for clarity. The index j and J loops over all the protein atoms and CG beads respectively. The definition of $f_j^{excl}(q)$ is shown in equation (2.2). On the other hand, $F_J^{CG}(q)$ is the CG form factor of the J^{th} CG bead, which is to be optimized in our method. $\mathbf{r}_j$ and $\mathbf{R}_J^{CG}$ are the atomistic and CG protein coordinates in corresponding representation.

Our fitting target is the scattering amplitude $A(\mathbf{q})$ for the entire continuous reciprocal space. However, this is infeasible in real calculations. Thus, the scattering amplitudes of a discretized set of $\mathbf{q}$ points are chosen to perform the

fitting only. For each scalar value of q , $M = 1000$ directions were selected uniformly on the q sphere, using the same spiral method to select directions in the spherical average method. The θ and φ angles representing the spherical coordinates of each $\mathbf{q}$ direction are determined by equation (2.5).

The Fourier transforms at different scattering vectors computed from CG form factors should be equal to those computed from the atomic representation in ideal case. As the form factors depend on the scalar value of q only, the following over-determined linear equations are then constructed for each q value by combining the above two equations (4.1) and (4.2) for all the directions with the same value of q .

$$\begin{pmatrix} \sum_J^{N_1} \exp(-i\mathbf{q}_1 \cdot \mathbf{R}_J) & \cdots & \sum_J^{N_1} \exp(-i\mathbf{q}_1 \cdot \mathbf{R}_J) \\ \vdots & \ddots & \vdots \\ \sum_J^{N_t} \exp(-i\mathbf{q}_M \cdot \mathbf{R}_J) & \cdots & \sum_J^{N_t} \exp(-i\mathbf{q}_M \cdot \mathbf{R}_J) \end{pmatrix} \begin{pmatrix} F_1^{CG}(q) \\ \vdots \\ F_t^{CG}(q) \end{pmatrix} = \begin{pmatrix} \sum_j^{N_a} f_j^{excl}(q) \exp(-i\mathbf{q}_1 \cdot \mathbf{r}_j) \\ \vdots \\ \sum_j^{N_a} f_j^{excl}(q) \exp(-i\mathbf{q}_M \cdot \mathbf{r}_j) \end{pmatrix} \quad (4.3)$$

In the above equation, t is the number of CG bead types in the CG representation of protein, where $t = 20$ for one-bead per-residue models and $t = 21$ for two-bead per-residue models. N_1 to N_t denote the number of CG beads in the protein with CG bead type 1 to t . N_a is the number of atoms in the atomistic protein structure. $\mathbf{r}_j$ and $\mathbf{R}_j^{CG}$ are the atomistic and CG coordinates as shown above. $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_M$ denote the $M = 1000$ scattering vectors whose directions are uniformly distributed on a sphere of radius q , as determined by the spiral method. In our computation, the value of scalar q is selected to be from 0 to $0.5 \text{ \AA}^{-1}$ with an interval of $0.01 \text{ \AA}^{-1}$. Theoretically, our EDM method can perform optimization of CG form factors up to arbitrary q range. In reality, the upper

limit of q (q_{max}) in the optimization process should be determined by the resolution of CG scheme used. Using a very high q_{max} value for CG representation of protein structure usually results in over-fitting.

The above equation (4.3) is only for single protein structures. However, it is straight forward to include multiple protein structures in the optimization of CG form factors. This can be done by simply applying equation (4.3) to all the structures in our dataset and combine all the equations together, which is shown as follows.

$$\begin{aligned} \mathbf{A}_1 \mathbf{f} &= \mathbf{b}_1 \\ \mathbf{A}_2 \mathbf{f} &= \mathbf{b}_2 \\ &\dots \\ \mathbf{A}_p \mathbf{f} &= \mathbf{b}_p \end{aligned} \quad \text{or } \mathbf{A} \cdot \mathbf{f} = \mathbf{b} \quad (4.4)$$

Here, P is the number of protein structures in our dataset. Each row in the left corresponds to the simultaneous equations for each protein structure as shown in equation (4.3). In the right part of the equation, Matrix $\mathbf{A}$ and $\mathbf{b}$ are constructed by $\mathbf{A} = (\mathbf{A}_1^T \quad \mathbf{A}_2^T \quad \dots \quad \mathbf{A}_p^T)^T$ and $\mathbf{b} = (\mathbf{b}_1^T \quad \mathbf{b}_2^T \quad \dots \quad \mathbf{b}_p^T)^T$. Because equation (4.4) is an over-determined equation, it is nearly impossible to construct a set of CG form factors that can fully satisfy all the equations in (4.4). Instead, the equation is solved by the least square fitting method in order to obtain an optimal solution that minimizes the deviation between the two sides of the equation. Equation is rewritten in the form of normal equation defined as follows.

$$\mathbf{A}^T \mathbf{A} \cdot \mathbf{f} = \mathbf{A}^T \mathbf{b}, \text{ or } \mathbf{G} \cdot \mathbf{f} = \mathbf{d} \quad (4.5)$$

$\mathbf{A}^T$ means the transformation of matrix $\mathbf{A}$. The matrix $\mathbf{G}$ is the normal matrix of the original matrix $\mathbf{A}$, while the vector $\mathbf{d}$ equals to $\mathbf{A}^T \mathbf{b}$. Equation (4.5) can be solved by a lot of numerical approaches. Here, the singular value decomposition (SVD) method [64] is used to find the solution. The resulting CG form factors represent the best fitting of the Fourier transforms of the atomistic electron density distributions, as guaranteed by the least square fitting algorithm.

4.1.2 Isolated bead approximation (IBA) method

We also implemented the isolated bead approximation (IBA) method, which is widely applied in previous works, in order to make a comparison with our EDM method. In the IBA approach, the CG form factor is evaluated by regarding all the atoms contained in the CG bead as a small isolated system. The square root of the scattering intensity of these atoms are computed and averaged over all possible conformations existed in the selected protein structure dataset, resulting in the final CG form factor. The formula to perform the computation is shown in the following.

$$F^{CG}(q) = \left\langle \left[\sum_{i,j}^{N^{CG}} f_i^{excl}(q) f_j^{excl}(q) \frac{\sin(qr_{ij})}{qr_{ij}} \right]^{1/2} \right\rangle_{PDB} \quad (4.6)$$

Here, i and j loop over all the atoms within the certain CG bead and the number of atoms is given by N^{CG} . The ensemble average is performed over all the possible conformations of the corresponding atom group in our dataset containing 1003 protein structures.

The square root in equation (4.6) will only generate positive form factor values. This is generally true for the one-bead coarse-grained scheme as used in the original work of the IBA method. However, for less coarse-grained case, there can be negative bead form factors because the excluded volume effect is considered here, as the atomic form factors we are using is $f^{excl}(q)$ calculated by equation (2.2). The modified atomic form factor $f^{excl}(q)$ of the hydrogen atom becomes negative near $q = 0$, with the value at $q = 0$ is -0.72. Theoretically, the CG bead form factor at $q = 0$ should be exactly the same as the summation of all the atomic form factors within the CG bead, as shown by the following equation.

$$F^{CG}(q = 0) = \sum_j^{N^{CG}} f_j(q = 0) \quad (4.7)$$

For those CG beads that contain many hydrogen atoms, their form factors at $q = 0$ could be negative. As discussed in the methodology part, two CG schemes at different CG levels are used. It is found that negative $F^{CG}(q = 0)$ values will present for several CG beads representing the residue side-chains, such as alanine, leucine and valine, in the two-bead per-residue scheme. Thus we need to perform corrections to CG form factors that do not satisfy equation (4.7), following the method used by Niebling et al. [41]. A sample correction process is shown in Figure 4.1. The red curve denotes the original CG form factor of the CG bead representing the side-chain of alanine. The values at small q region are wrongly computed to be positive by equation (4.6). The data points after the inflection points (blue stars) should be correct form factor values. Thus they are chosen to fit a six-order polynomial, using the $F^{CG}(q = 0)$ computed from equation (4.7) as fitting constraint (green sphere). The CG form factors are then recalculated using the resulting polynomial, which results the green curve.

Figure 4.1 Correction of CG form factors obtained by the IBA method. The red curve is the original CG form factors of CG bead representing alanine side chain. The positive values at low q region are incorrect. So we use only the values after the reflection point, shown in blue stars, to perform a six-order polynomial fitting in order to obtain reasonable values for the CG form factors at low q region. The value of $F^{CG}(q = 0)$ evaluated from the summation of all the atomic form factors in the CG bead, shown by the green sphere, is used as fitting constraint. The resulting new CG form factors are shown in green curve.

4.2 Method to model experimental SAXS profiles

4.2.1 Determine the contribution of protein hydration layer

The EDM-derived CG form factors are fitted to reproduce the scattering of protein molecules only. Thus in order to validate them using experimental SAXS profiles, we need also to include the contribution of the protein hydration layer, which is represented by explicit water molecules in this study. The coordinates of the water molecules are determined by a stringent method described as follows.

A pre-equilibrated cubic water box was generated from constant NPT MD simulations of 1 ns at the condition of 300 K and 1 bar. The SPC/E water model [65] was applied in the simulation and the box size after equilibrium was 9.9852 nm. The protein structure was then overlapped with the pre-equilibrated water box in order to mimic the solvated protein system. If the size of the protein is too large, the box size was extended by simply adding periodic images of the water box along necessary dimensions. The following criterions are used to determine whether a water molecule is inside the protein hydration layer or not. A certain water molecule was accepted to be part of the protein hydration layer if the minimum distance between the water oxygen atom and its nearest protein atom was larger than $r_{vdw} + r_{wat}$ and smaller than $r_{vdw} + r_{wat} + 3 \text{ \AA}$. In the expression, 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 water molecule taken to be $1.4 \text{ \AA}$. The water molecule is considered to be partially overlapped with protein atoms if the computed minimum distance was smaller than $r_{vdw} + r_{wat}$ but larger than r_{vdw} . In this case, a stochastic procedure is performed, following the work by Grishaev et al. [31], to decide whether or not to accept this water to be part of the protein hydration layer. The probability of acceptance is computed using the following formula.

$$P_{accept} = \prod_j \left(\frac{r_j - r_{vdw}^j}{r_{wat}} \right)^\gamma \quad (4.8)$$

Here, j loops over all the protein surface atoms whose distances (r_j) to the oxygen atom of the specific water molecule are smaller than $r_{vdw} + r_{wat}$. r_{vdw}^j is the van der Waals radius of the j^{th} protein atom. The empirical parameter γ is set to be 0.1, following the value used by Grishaev et al. [31]. The value has been chosen to best reproduce the excluded volume of the protein molecule. The

physical meaning of this formula is that if a specific water molecule overlaps with multiple protein atoms, it is more likely to be displaced water instead of hydration water. A larger acceptance ratio will result in underestimating of the excess hydration weight, while a smaller acceptance ratio will overestimate the excess hydration weight. In order to eliminate those water molecules that posited in the small cavities with the protein molecule, an additional post-processing step is performed in the end. Basically, the water molecules owning less than two neighboring water molecules within distance of 3.8 Å are removed from the hydration water set. This procedure is performed iteratively until no more water molecules are removed further more.

During the computation of CG SAXS profiles, the hydration water molecules are included also in CG form, with the CG beads placed at oxygen atoms of the molecules. The isolated bead approximation is good enough for CG water molecules as the water molecule is naturally isolated from each other. Thus the CG form factor of water molecule is simply obtained using the following equation.

$$F_w^{CG}(q) = \left[\sum_{i,j}^3 f_i(q) f_j(q) \frac{\sin(qr_{ij})}{qr_{ij}} \right]^{1/2} \quad (4.9)$$

The distances r_{ij} between atoms in the water molecules are taken from the standard SPC/E water model. Following the work by Sorenson et al. [66], the atomic form factors of the oxygen and hydrogen atom are modified in order to consider the electron withdraw effect. The contribution of hydration water molecules should be scaled by a small weight factor w_{wat} , because only the excess electron density of hydration layer compared with bulk water contributes

to the final scattering intensity. This is done by scaling the CG form factor of water, i.e., $F_w^{CG'}(q) = w_{wat} F_w^{CG}(q)$.

4.2.2 Scalable excluded volume

The Gaussian sphere approximation is used to compute the contribution of the protein excluded volume during our fitting of CG form factors. However, the summation of the van der Waals volume of the protein atoms sometimes cannot accurately reproduce the real protein excluded volume. Thus, in order to obtain better fitting with experimental scattering intensities, many computational tools have one free parameter that can tune the exclude volume term. Similarly, we also try to make the contribution of the protein exclude volume tunable in our work. This is done by decomposing the CG form factor obtained by our EDM method into two terms, one counts for the scattering of protein atom in vacuum only and another represents the contribution of the excluded volume effect. After performing the EDM optimization described previously, we reconstruct the over-determined simultaneous equations (4.4) applying the original atomic form factors that do not consider the excluded volume effect, as shown in equation (2.1). The resulting CG form factors by solving this equation using the least square fitting approach include the scattering of protein atoms in vacuum only and are denoted to be F_{vacuum}^{CG} . Then the other decomposed term that is responsible for the protein excluded volume effect can be calculated by subtracting the overall CG form factors obtained with F_{vacuum}^{CG} straightforwardly, which is shown by the expression $F_{excl}^{CG} = F_{vacuum}^{CG} - F^{CG}$. Finally, a modified set of CG form factors with the contribution of the excluded volume term tune can be expressed by $F^{CG'} = F_{vacuum}^{CG} - w_{excl} F_{excl}^{CG}$, where w_{excl} is

the weight factor. It is shown that better fitting to the experimental scattering intensities can be obtained by tuning the free parameter w_{excl} .

4.2.3 Optimization of weight factors

As discussed in the previous two parts, the two weight factors w_{wat} and w_{excl} can be optimized in order to better reproduce the experimental SAXS profiles, when the CG form factors from our EDM method is used together with explicit hydration water to model the experimental data. The target of optimization is the χ^2 value between theoretically computed and experimental SAXS profiles. As the explicit hydration water is used, the analytical derivatives of the target χ^2 value to the two free parameters are hardly available. Thus the Nelder-Mead simplex algorithm [67], which is non-derivative based, is implemented to perform the optimization of the two free parameters. We choose the initial value to be $w_{wat} = 0.03$ and $w_{excl} = 1.0$. The convergence criterion is set to be the size of the simplex smaller than 10^{-3} .

4.3 Amino acid form factors

In our study, there are totally four CG schemes at two different coarse-grained levels. The CG form factors of each CG scheme are evaluated using both our EDM method and the IBA method. This results in six sets of CG form factors, as the results of the IBA method are same for the different bead positions at the same coarse-grained level. The resulting form factors and the coarse-grained schemes are shown in Figure 4.2, Figure 4.3 and Figure 4.4.

Figure 4.2 The coarse-grained scheme and CG form factors obtained from electron density matching (EDM) and single bead approximation (IBA), respectively. (a) Illustration of a two-bead-per-residue CG scheme where each bead is placed at its center of electron (COE). For the illustrative structure shown in sticks, the backbone and side-chain of each amino acid is represented by a blue and red bead, respectively. (b) and (c) Resulting CG form factors from EDM and IBA, respectively, where the blue and red curves represent form factors of backbone and side-chain beads of all 20 residues, respectively. The ordering of their side-chain form factors is following 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.

Figure 4.3 The CG form factors obtained by (b) EDM and (c) IBA, using the one-bead COE CG scheme illustrated in (a). The red curves show the CG form factors representing the 20 types of protein residues. The ordering of the residue form factors is the 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.

Figure 4.4 The CG form factors obtained by the EDM method using (a) the two-bead $\text{Ca} + \text{COM}$ and (b) one-bead Ca CG schemes respectively. The blue curve in (c) shows the CG form factor of the residue backbone. The red curves in (c) represent the CG form factors of the 20 types of residue side-chains, while the red curves in (d) represent the CG form factors of the 20 types of protein residues. The ordering of the side-chain form factors in (c) and the residue form factors in (d) is the 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.

In Figure 4.2, the CG form factors are shown for the two-bead per-residue scheme with the COE used to determine the position of both backbone and side-chain beads. The coarse-grained scheme is illustrated in Figure 4.2(a), while the CG form factors results of our EDM method and the IBA method are plotted in Figure 4.2(b) and Figure 4.2(c) respectively. It is seen from the figure that the resulting CG form factors of the backbone and side-chains have quite different trend. This is not surprising due to the relatively large variations in side-chains CG beads. For side-chain beads, the results of our EDM method generally show a sharp increment in the small q range of $0 \sim 0.3 \text{ \AA}^{-1}$. Some of them reach their maximum value at around $q \approx 0.3 \text{ \AA}^{-1}$ and then begin to decrease after this point. In contrast, the CG form factors derived by the IBA method generally show a relatively slow but continuous increment over the entire q range of $0 \sim 0.5 \text{ \AA}^{-1}$. Their maximum values usually emerge at $q > 0.5 \text{ \AA}^{-1}$ (data not shown). Besides, for the CG form factors of the backbone, our EDM result has a local minimum at q value around $0.3 \text{ \AA}^{-1}$, while the IBA result remains almost unchanged for the entire q range.

The observed differences between the resulting CG form factors of our EDM method and the IBA method should be mainly caused by the different

treatments of the correlations between the orientations of different CG beads. In the IBA method, CG beads are supposed to be “isolated” with each other. This means the CG beads can freely rotate and are not biased to certain orientations in different local environments. No correlations between different CG beads are taken into consideration when computing the form factors. However, in our EDM method, the fitting target is the Fourier transform of the entire protein electron density, which is related to the CG form factor of every bead. When computing the normal matrix $\mathbf{G}$ in equation (4.5) by $\mathbf{A}^T \cdot \mathbf{A}$, the structure information of different types of CG beads are correlated with each other. Essentially, the correlations between different types of CG beads are explicitly considered by the off-diagonal elements of the matrix $\mathbf{G}$.

The CG form factors for the two method using the one-bead COE CG scheme are shown in Figure 4.3 and the CG form factors for the EDM method using two-bead $C\alpha$ +COM and one-bead $C\alpha$ scheme are plotted in Figure 4.4. It can be seen from the figures that the EDM results are quite different from those of the IBA method in all the mapping schemes. Additionally, the optimized form factors are also different from those obtained from the work by Stovgaard et al. [42], where the non-linear fitting of atomistic scattering profiles are used to perform the optimization. The different fitting targets and numerical optimization schemes may be the main reason of the observed discrepancies.

4.4 Accuracy and efficiency of CG form factors

The accuracy of CG form factors derived above is evaluated by comparing the CG and atomistic SAXS profiles, computed by the CG form factors and the atomic form factors respectively, for all the protein structures in our dataset.

Smaller discrepancies between the CG and atomistic SAXS profiles indicate the set of CG form factors tend to be more accurate. In this section, we compared the accuracy of the EDM-derived CG form factors with the IBA method as well as the CG form factors reported in the work of Stovgaard et al. [42].

As the first step, the accuracy of CG form factors derived from our EDM method is compared with the IBA method. The R value defined in equation (2.11) is used to quantify the difference of the CG and atomistic SAXS profiles for a single protein structure. The distribution of R values is then evaluated from our protein structure dataset and used to compare the accuracy between different methods. The R value distributions of all the 1003 high-resolution protein structures using different CG schemes and the EDM/IBA method are plotted in Figure 4.5. The results of two-bead CG schemes are shown in (a), while the one-bead case shown in (b). The CG SAXS profiles computed from our EDM method are found to be much more accurate than those from the IBA method from the comparison between the EDM distributions (shown in red) and the IBA distributions (shown in blue) in both figures. Figure 4.5(a) shows 97.2% and 91.2% of all the R values in the distributions are smaller than 0.1 for our EDM method when two-bead COE and two-bead $C\alpha$ +COM schemes are used respectively. On the other hand, only 6.08% and 5.58% of all the R values are smaller than 0.1 for the IBA method. Moreover, there are 16.8% and 4.6% of R values whose values are larger than 0.4 for the two schemes respectively. When the one-bead CG schemes are used, the R values are generally larger than the relatively finer two-bead schemes as shown in Figure 4.5(b). However, the similar trend that our EDM method outperforms the IBA method is observed as well.

Figure 4.5 Distribution of R values describing the difference of SAXS profiles between CG and all-atom models. The SAXS profiles within q range of $0 \sim 0.5 \text{ \AA}^{-1}$ are used in the computation. Two different CG schemes are shown: (a) two-bead-per-residue model and (b) one-bead-per-residue model. The results of EDM and IBA are presented in red and blue curves, respectively, where spherical circles represent that the position of each CG bead is placed at its corresponding COE and cubic points represent that the position of each bead is placed at its C α atom (in the case of a one-bead-per-residue model) or at the C α atom for its backbone and the center of mass (COM) of its side-chain (in the case of a two-bead-per-residue model).

Figure 4.6 Distribution of R values describing the difference of SAXS profiles between CG and all-atom models. Two different CG schemes are shown: (a) two-bead-per-residue model and (b) one-bead-per-residue model. The results of EDM and IBA are presented in red and blue curves, respectively, where spherical circles represent that the position of each CG bead is placed at its corresponding COE and cubic points represent that the position of each bead is placed at its C α atom (in the case of a one-bead-per-residue model) or at the C α atom for its backbone and the center of mass (COM) of its side-chain (in the case of a two-bead-per-residue model). The R values are computed for the q range of 0 ~ 0.35 $\text{\AA}^{-1}$.

In some SAXS experiments, only the scattering intensity of region $q \leq 0.35$ $\text{\AA}^{-1}$ is reported. Thus we also computed the distributions of the R values based

on the comparison of this low q region only. The results are illustrated in Figure 4.6. It can be seen that the CG form factors derived from our EDM method also have a much better accuracy than the IBA method for the relatively low q region.

The more accurate performance of the EDM-derived CG form factors originate from the consideration of the correlations between different coarse-grained beads, especially the orientational correlations. For example, chemical bonds exist between nearby backbone atom groups as well as the backbone and side-chain of the same residue, thus their orientations should be relatively fixed. Similarly, nearby side-chain atom groups can also interact with each other, resulting in some preferences for their relative orientations. These orientational heterogeneities are taken into consideration in our EDM-derived CG form factors which have improved accuracy. On the other hand, the Debye formula used by the IBA method implicitly assumes the atom groups corresponding to the CG beads can rotate freely, ignoring the orientational heterogeneities discussed above. This kind of assumption works very well for CG water molecules and it is only reasonable for very low- q regions, where the molecular details are not required when applied to protein molecules.

From Figure 4.5 and Figure 4.6, the different mapping schemes at the same CG level can also result in slightly different R value distributions. Thus it is interesting to find out the better mapping scheme for each method to parameterize the CG form factors. For our EDM method, the CG form factors using the COE mapping scheme always have better performances in both two-bead and one-bead per-residue cases. In contrast, for the IBA method, the COE type mapping schemes seem to perform worse. The R value distribution of the two-bead $C\alpha$ +COM scheme is only slightly better than the two-bead

COE scheme, while that of the one-bead C α scheme is obviously more accurate compared with the one-bead COE scheme. It is not surprised that the COE mapping scheme is a better choice when our EDM method is used, because of the fitting target of the EDM method is equivalent to the electron density distribution of the protein molecule. However, for the IBA method, it is hard to determine which bead location is a better choice although the C α or C α +COM mapping scheme have better R value distributions. The inspections of the fitting curves show that the deviations of the CG and atomistic SAXS profiles mainly occur at q region around $0.3 \text{ \AA}^{-1}$ for the C α type mapping schemes, while the deviations appear in larger q values for the COE type mapping schemes. As can be seen in Figure 4.6(b), the result of one-bead COE is somewhat better than the one-bead C α CG scheme when the region $q \leq 0.35 \text{ \AA}^{-1}$ is used for R value evaluation.

Figure 4.7 Average R values between CG and all-atom SAXS profiles as a function of q_{cut} values. The R values were evaluated over the q -range below q_{cut} and averaged over all the 1003 protein structures. The EDM method was used for the derivation of the CG form factors. Results from both one-bead and two-bead COE CG schemes are shown by blue and red curves, respectively.

The applicable q -ranges of coarse-grained schemes at different levels are also important in real applications. Different q_{cut} values were used as maximum q values in the computation of the R values for the protein structures in our dataset. The EDM derived CG form factors of both one-bead and two-bead COE CG schemes are used in our test and the average R values, which indicates the error scale of the corresponding CG form factors, are plotted in Figure 4.7. From the figure, it is seen that both curves are increasing as larger q_{cut} values are used for the computation of R value, as expected. It is found that if the average R value of 0.1 is set to be acceptable error value, then the CG form factors of one-bead COE model show reasonable performance for a q_{cut} value of as large as $0.35 \text{ \AA}^{-1}$. For the CG form factors parameterized by IBA method, the reasonable q_{cut} value is smaller compared with the EDM method. We conclude that the one-bead per-residue CG model is only useful when computing the scattering profiles for relatively low- q regions. In order to accurately reproduce the scattering intensities for larger q -ranges, two-bead per-residue or models with higher resolutions are necessary.

Figure 4.8 Distribution of R values describing the difference of SAXS profiles between CG and all-atom models for a dataset containing NMR structure ensembles of intrinsic disordered proteins. Six proteins (PDB entries: 1A5R, 1R8U, 2FFT, 2L14, 2LJL, and 2LY4) with total 100 structure models are contained in the dataset. The form factors obtained by our EDM method together with the two-bead COE CG scheme are used to compute the CG SAXS profiles.

Table 4.1 Comparison of the mean and standard deviation of R values for the six intrinsic disordered protein NMR structure ensembles, using our EDM method. Results from one and two bead COE models are shown.

PDB (No. of structures)	R (2-bead COE)	R (1-bead COE)
1A5R (10)	0.057 ± 0.022	0.115 ± 0.054
1R8U (20)	0.028 ± 0.010	0.425 ± 0.076
2FFT (20)	0.017 ± 0.007	0.090 ± 0.037
2L14 (20)	0.026 ± 0.013	0.183 ± 0.036
2LJL (20)	0.058 ± 0.018	0.174 ± 0.060
2LY4 (10)	0.049 ± 0.022	0.144 ± 0.047

One important application of the solution X-ray scattering is to study the structure ensembles of intrinsically disordered proteins (IDPs), usually together with NMR experiments. In our study, the CG form factors are parameterized using well-structured protein molecules from our dataset. Thus it is of great interest to check whether these CG form factors will work for IDPs or not. Six IDP structure ensembles obtained from NMR experiments, which contain 100 structure models in total, are selected to be our test cases. The PDB entries of the six IDPs are: 1A5R [68], 1R8U [69], 2FFT [70], 2L14 [71], 2LJL [72] and 2LY4[73] and the radii of gyration of all the structure models are within the measurement of SAXS experiment, ranging from 1.4 to 4.3 nm. Similar to previous results, the distribution of R values obtained by the two-bead COE scheme is shown in Figure 4.8. It can be seen the R values are less than 0.1 for all the 100 NMR structure models, and the peak value of the histogram locates at the region of $R = 0.02 \sim 0.03$. This indicates the CG SAXS profiles evaluated from our EDM-derived form factors for two-bead COE model can well reproduce that from atomic form factors. Additionally, the mean and standard error of R values for each IDP are listed in Table 4.1 for both two-bead and one-bead COE CG schemes. The result of the one-bead model can also reasonably reproduce the atomic SAXS profiles, although the R values of the one-bead model are relatively larger compared with those of the two-bead model. Overall, the EDM-derived CG form factors are shown to have a good performance for the disordered protein structures despite being parameterized using protein native structures.

Figure 4.9 Comparison of the S value distribution of the EDM method with that reported by Stovgaard et al. [42]. S is evaluated by

$$S = \sqrt{\frac{1}{N_q - 1} \sum_q \left[\frac{I_{CG}(q) - I_{AA}(q)}{\sigma(q)} \right]^2}, \text{ where the empirical standard error}$$

$\sigma(q_i) = I(q_i)(q_i + \alpha)\beta$ with $\alpha=0.15$ and $\beta=0.3$. We use the same protein dataset from Stovgaard et al. The red histogram shows the distribution of S values from our method, and the blue histogram shows the results reported by Stovgaard et al..

Furthermore, the accuracy of our EDM-derived form factors is compared with that obtained from non-linear fitting reported by Stovgaard et al. [42]. The maximum likelihood method is used to perform the fitting, which targets at the scattering intensities evaluated from Crysol without considering the contribution of the hydration layer. The comparison is carried out using the small validation dataset containing 50 protein structures from their work, for which the S values are reported. The so-called S value used as comparison criterion is roughly the square root of the R value defined in our study. Figure 4.9 shows the distribution of S values from our EDM method compared with those reported by Stovgaard et al.. It can be seen that the EDM-derived CG

form factors perform slightly better. Nevertheless, both methods can accurately reproduce the scattering intensities obtained from Crysol for all the testing protein structures.

The improvement of computational efficiency by using the CG form factors was benchmarked using a huge protein complex GroEL:GroES2 (PDB entry: 4PKN) [74], containing 8689 residues and 130486 atoms in total. The spherical average algorithm was used to perform orientational average in the calculation of scattering intensities. Computation based on all-atom protein structures took 203 seconds, while only 25 seconds were required when the two-bead CG scheme was used. The above computational time was obtained on single Intel Xeon E7-8850 CPU at 2.00 GHz frequency using our own implementation of the spherical average method. The observed eight fold increment of computational efficiency is roughly proportional to the total number of beads in the all-atom and CG case. Moreover, the computational time can be further reduced with a slightly trade-off of the accuracy, by using the one-bead per-residue models.

We consider only the scattering intensity of the protein part in the benchmark case shown above. The SAXS modeling also need to include the contribution of hydration water molecules if the results are to be compared with experimental data. In this study, the explicit hydration water molecules are determined by superimposing the protein molecule with pre-equilibrated water box, as described in previous method part. The determination of water molecule positions and the scattering amplitude of the explicit water molecules will also take significant computational time, especially for relatively large protein system. However, there are many other computational methods to treat the

hydration layers as shown in literature. In principle, implicit treatments like the one used by Crysol, can also be applied to our EDM-derived CG form factors.

When performing structure modeling from SAXS profiles, the scattering intensities are usually computed for many times. In this case, the improvement of computational efficiency using CG models can be significant. Recently, there are also computational tools that use CG molecular representation to perform SAXS profile-based structure predictions, such as SCT by Wright et al. [75]. Additional steps are needed to convert the CG representation back to atomistic structure. The extra computational time required by the model conversion is nearly neglectable as it is only done for once during the whole calculation. Software to perform this kind of backward mapping is readily available in literature for both the C α -based and arbitrary mapping [76, 77]. Moreover, the CG form factors can be used in SAXS-guided MD simulations to sample the structure ensembles that correspond to the experimental SAXS profiles. The advantage of using CG representation is also reflected by speeding up of molecular dynamics simulation. Considering the two-bead COE coarse-grained scheme, which is shown to have the best fitting to atomistic scattering intensities, we can use simple structure-based potentials to perform conformational restraints when performing structure modeling, similar to that done by Zheng et al. [25]. Furthermore, standard multiscale coarse-graining methods [78-81] or statistical approaches [82-85] can be used to parameterize the CG force fields in order to conduct SAXS-guided CG MD simulations with our EDM-derived CG form factors.

Figure 4.10 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 in a red dotted line. CG results using the EDM method (COE), the IBA method (COE) and the IBA method ($C\alpha$ + 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.

Figure 4.11 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 in 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.

4.5 Example scattering profiles

In order to visualize how well the CG SAXS profiles computed using the derived CG form factors can reproduce the atomistic SAXS profiles, we plotted the SAXS profiles for four selected example proteins with different sizes in Figure 4.10 and Figure 4.11. The PDB entries and number of residues of the four proteins are 4CO8 (98 residues) [86], 4JL5 (203 residues) [87], 4UYB

(400 residues) (To be published) and 1N62 (804 residues) [88]. The result of CG SAXS profiles computed using the two-bead per-residue schemes are shown in Figure 4.10, while those of the one-bead per-residue schemes are illustrated in Figure 4.11. From both figures, the CG SAXS profiles computed using our EDM derived form factors can better reproduce the atomistic ones better than the IBA method. Taken the two proteins shown in Figure 4.10(a) and Figure 4.10(b) as examples, the scattering intensities evaluated from the IBA-derived form factors with two-bead COE CG scheme show significant discrepancies compared with the atomistic profiles in the intermediate q -range of $0.2 \sim 0.3 \text{ \AA}^{-1}$. On the other hand, the results from the EDM-derived form factors using the same CG scheme overlap with their all-atom counterparts within line width for all the four example proteins. It is also noticed that three out of the four results of the two-bead $C\alpha$ +COM mapping scheme are better than the two-bead COE scheme for the IBA method. This observation is consistent with our previous discussion. For the one-bead CG models, the scattering intensities from the IBA show even larger deviations, while similar conclusions still persist for the comparison between the IBA and our EDM method, as shown in Figure 4.11. For the one-bead $C\alpha$ model, there are two proteins (1COB and 4JL5) whose main deviations appear at the intermediated q region of $0.2 \sim 0.3 \text{ \AA}^{-1}$. Compared with the one-bead COM CG schemes, where the major discrepancies occurs at relatively larger q region, the results of $C\alpha$ model will be better for the whole q range, while the COE CG scheme becomes better if we only fit the low q part ($q \leq 0.35 \text{ \AA}^{-1}$). This can also be validated by the distribution of the R values shown in Figure 4.5 and Figure 4.6. In conclusion, the CG form factors obtained by our EDM method perform better than those of the IBA method. While the IBA method still has reasonable

accuracy for the less coarse-grained models, the more rigorous EDM method is necessary if more coarse-grained structure representations are used.

4.6 Dependency of our EDM method to the protein dataset

We have used exactly the same protein structure dataset to perform both the parameterization and the validation of the CG form factors using our EDM method in the previous sections. There are concerns that the good performance shown above is the result of over-fitting. In order to verify the reliability of our EDM method, including the transferability as well as the convergence, the CG form factors are parameterized using different training sets of protein structures and compared with each other.

As the first step, the whole protein structure dataset containing 1003 structures was randomly split into two subsets containing 503 and 500 structures respectively. Using the two-bead COE CG scheme, we optimized two sets of CG form factors using the two subsets as training set respectively. The discrepancy of the two sets of CG form factors are measured by difference

value defined as $d = \frac{1}{N_{CG}} \sum_{CG \text{ type}} \left[\frac{1}{N_q} \sum_q (f(q) - f_0(q))^2 \right]$. Here, N_{CG} is the

number of CG bead types, which equals to 21 for the CG scheme used in this part, and N_q is the number of data points on each CG form factor curve, which is 51 for q ranging from 0 ~ 0.5 Å⁻¹ with interval of 0.01 Å⁻¹. The difference value d is evaluated to be only 5.49×10^{-3} for the two sets of CG form factors optimized from two isolated protein structure subsets. This absolute difference value corresponds to a relative error of around 1.5%, because the CG form

factor values are generally within the range of $-5 \sim 15$. Thus the obtained two sets of CG form factors are nearly identical.

Figure 4.12 Comparison of CG SAXS profiles computed from different sets of CG form factors. The red line with spheres shows the CG SAXS profiles evaluated with the CG form factors obtained from protein set 1, while the green line shows the results when protein set 2 is used for parameterization. The definition of the two protein sets can be found in section 3.4 of the main text. The blues lines show the relative errors using the atomistic results as the standard. The two proteins are (a) Lysozyme (1LYZ) with 129 residues and (b) Ubiquitin (1UBQ) with 76 residues.

Figure 4.13 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 dataset size used for the derivation of CG form factors. The data points show the normalized differences between the form factors from a protein dataset of N_{PDB} structures and those from the largest protein dataset of 1003 structures. The range of N_{PDB} is from 100 to 1003. The definition of the normalized difference can be found in main text.

The similarity of the two sets of CG form factors is further verified by comparing the computed CG scattering intensities for two example proteins: Lysozyme (PDB ID: 1LYZ) [59] and Ubiquitin (PDB ID: 1UBQ) [89]. The resulting CG SAXS profiles and the differences between the results from different CG form factors are shown in Figure 4.12. The discrepancies of the CG scattering intensities are shown to be within the line thickness for both proteins and the maximum relative differences values between the CG scattering intensities are only 6‰ for lysozyme and 3‰ for ubiquitin, as illustrated in the bottom panel of each figure. This small difference between the resulting CG SAXS profiles can be ignored in real applications.

In order to determine whether there exists over-fitting issue in our EDM method, the distributions of R values are also evaluated using the alternative protein subset from that used to parameterize the CG form factors. As shown in Figure 4.13(a), the R values of both distributions are very small, with almost all of them below 0.1, and the two distributions only slightly different from each other. Overall, the tests here suggest that the CG form factors obtained in previous section do not rely on the selection of the protein structure dataset. The high accuracy of our EDM-derived CG form factors is transferable to different proteins other than those used for parameterization.

We then investigated the convergence of our EDM method with the increasing of protein dataset size. The CG form factors were obtained from different number of protein structures, varying from 100 to 1000, and compared with the one optimized using all the 1003 structures. The difference value d is used to measure the difference between different sets of CG form factors. As seen in Figure 4.13(b), the resulting d values decrease from 1.33×10^{-2} to 2.14×10^{-6} when the number of protein structures used to perform the optimization

increases from 100 to 900 and it is found the reduction of d value is significant only before the size of 300, where it reaches a relatively small value of 1.98×10^{-3} . We conclude that our EDM-derived CG form factors used above have reached a converged value by using a dataset with size much larger than 300 for parameterization.

4.7 Accurate reproduction of experimental scattering profiles

In order to perform structure modeling with the derived CG form factors, it is important to make sure that the CG SAXS profiles obtained are comparable with the experimental data. Two solvent effects will significantly affect the resulting comparisons. One is the hydration layer and the other is the excluded volume of each protein molecule. The explicit representation of hydration layer is used in this study and the positions of the explicit hydration water molecules are determined by the procedure shown in the method part. As there may be errors when using the summation of van der Waals volumes of protein atoms to estimate the displaced solvent volume of the whole, this term is also fine-tuned during the fitting to experimental SAXS profiles. Two free parameters counting for the above two effects are implemented in many methods, such as Crysol, in order to reproduce the experimental profiles. We also adopted a similar approach here. The EDM-derived CG form factors were first decomposed into the vacuum term representing the scattering of protein atoms in vacuum and the excluded volume term representing the scattering of displaced solvent molecules. The weighting factors of the hydration layer and the protein displaced volume were optimized for best fitting to experimental SAXS profiles using the Nelder-Mead simplex algorithm [67].

Figure 4.14 Decomposition of the CG form factor representing the backbone of residues to a vacuum term and an excluded volume term. The red curve is the fitted CG form factor using atomic form factors f^{excl} shown in Eq. (2.2), which considers the excluded volume effect. The green curve is the CG form factor evaluated by using the original atomic form factors shown in Eq. (2.1), 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 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.

The decomposition of the CG form factor, using the backbone CG bead as an example, is shown in Figure 4.14. The original CG form factor (shown in red curve) is parameterized using the atomic form factors $f^{excl}(q)$ defined by equation (2.2), which consider the excluded volume effect by the Gaussian sphere approximation. The vacuum term (shown in green curve) is optimized using atomic form factors $f(q)$ defined by equation (2.1), without considering

the excluded volume effect. Finally, the contribution of the excluded volume effect (shown in blue curve) is obtained by calculating the difference between the two terms. This term can be scaled by weight factor w_{excl} during the fitting to experimental SAXS profiles.

The CG SAXS profiles evaluated using the EDM-derived CG form factors are then compared with their corresponding experimental profiles for six benchmark proteins: Lysozyme (PDB entry: 1LYZ) [59], Ubiquitin (PDB entry: 1UBQ) [89], Myoglobin (PDB entry: 1WLA) [60], RNase (PDB entry: 1C0B) [61], Cytochrome C (PDB entry: 1HRC) [90] and Chymotrypsinogen A (PDB entry: 2CGA) [91]. We obtained the experimental SAXS profiles of the six proteins from the Small Angle Scattering Biological Data Bank (SASBDB), with their entries listed in Table 4.2. The experimental data points with q values smaller than 0.0311, 0.0601, 0.0248, 0.0631, 0.0201 and 0.0448 $\text{\AA}^{-1}$ were discarded for the six experimental SAXS profiles respectively, due to either the Guinier plots of the scattering intensities are non-linear or the experimental data are too noisy in this region. As discussed in the method part, the two weight factors: w_{wat} representing the contribution of the excess electron density of the hydration layer compared with the bulk water, and w_{excl} representing the weight of the excluded volume are optimized using the Nelder-Mead Simplex algorithm in order to have the best fitting to experimental data. The initial value of the two parameters are $w_{wat} = 3\%$ and $w_{excl} = 1.0$. Similar fitting parameters are also available in software like Crysol. Thus we compared the resulting weight factors and the χ^2 values for our CG form factors with those from Crysol, as listed in Table 4.2. The fitting results of all the six proteins are shown in Figure 4.15 and Figure 4.16.

Table 4.2 The best-fit weighting parameters yielding optimal χ^2 values for experimental SAXS data of all the six example proteins, using EDM and CRY SOL, respectively. The weighting parameter w_{wat} is to account for the scattering contribution of excess electron density in a surrounding hydration layer, while w_{excl} is for the effect of protein excluded volume. The SASBDB entries of the experimental SAXS profiles are listed under the protein name.

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

Figure 4.15 Comparison of EDM-based and experimental SAXS profiles. Two example proteins, (a) lysozyme and (b) ubiquitin, are used where their experimental SAXS profiles (in red) are 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 computed and experimental SAXS profiles, the scattering contributions from the surrounding hydration layer and from the protein excluded volume are optimized.

Figure 4.16 Comparison of EDM-based and experimental SAXS profiles for four additional example proteins: (a) Myoglobin, (b) RNase, (c) Cytochrome C and (d) Chymotrypsinogen A. The experimental SAXS profiles (in red) are 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 computed and experimental SAXS profiles, the scattering contributions from the surrounding hydration layer and from the protein excluded volume are optimized.

From Figure 4.15(a), it is seen that the CG scattering intensities of Lysozyme (shown in blue curve) computed using our EDM-derived CG form factors can reproduce the experimental intensity (shown in red curve) very well. The χ^2 value is only 0.895, even slightly smaller than the result of Crysol ($\chi^2 = 1.404$). After the optimization of the two weight factors, it is found a relatively small hydration shell weight ($w_{wat} = 0.864\%$) is included in the computation, while a

relatively large excluded volume ($w_{excl} = 1.051$) is used. On the other hand, when fitting the experimental SAXS profiles using Crysol, the resulting excess weight of the hydration layer equals to 1.5%, and the resulting effective atomic radius is 0.166 nm, corresponding to weigh factor of excluded volume $w_{excl} = 1.025$ in our formula.

The result of protein Ubiquitin is plotted in Figure 4.16(b), where good agreement between the CG SAXS profile and experimental data is observed. The χ^2 value of our CG SAXS profile is 0.960, which is also slightly smaller than that obtained from Crysol ($\chi^2=1.766$), although the fitted curves from both methods overlap with the experimental profile reasonably. The two free parameters are optimized to be $w_{wat} = 0.956\%$ and $w_{excl} = 1.050$ using our CG form factors, while the values obtain by Crysol are $w_{wat} = 3\%$ and $w_{excl} = 1.074$.

For the rest four example proteins, we also obtained similar χ^2 values using our CG approach and Crysol, as listed in Table 4.2. This result further validates the reliability of our EDM-derived CG form factors in the modeling of experimental SAXS profiles. From the comparison of the two weight factors obtained by our method and Crysol, it is found that the values of w_{excl} are highly consistent with each other, while the w_{wat} values are somewhat different from each other. The similar w_{wat} values indicate that the excluded volume terms in both methods have common properties. However, the difference of w_{wat} values should originate from the different treatments of the protein hydration layer in the two methods.

4.8 Conclusions

We have presented a novel method, named electron density matching (EDM), to parameterize the form factors for protein coarse-grained beads. Compared with the widely applied isolated bead approximation (IBA) method in literature, the accuracy of the EDM-derived CG form factors are greatly improved by least square fitting of the Fourier transform of the protein electron density. Thus, the EDM-based method can reduce the computational cost of SAXS intensity calculation by using CG representation, while maintaining high accuracy that is comparable to the computational methods using atomistic representation. In our study, the CG form factors for both two-bead and one-bead CG schemes are optimized.

When including the contribution of protein hydration layer and the protein exclude volume correction, the optimized CG SAXS profiles evaluated using the EDM-derived CG form factors can well reproduce the experimental SAXS profiles for all the benchmark proteins. Our EDM-based method is of great potential to be used in modeling the protein structure based on experimental SAXS profiles. Other possible applications of our EDM-derived CG form factors include computing the scattering intensities of very large protein complex systems as well as those low-resolution structures from experimental methods such as cryo-electron microscopy. The EDM method can also be applied to other biomacromolecular system other than protein, such as DNA.

5 Implicit modeling of protein hydration shell in solution small-angle X-ray scattering using modified form factors

5.1 Method to parameterize the implicit hydration term

Scattering data usually contain insufficient information to solve the high-resolution structure of a protein. It is a common practice to construct a hybrid pseudo energy function in structure optimization including both structure-constraint terms and SAXS-fitting terms. MD simulations or direct minimizations based on such a hybrid pseudo energy function can determine the reasonable structures with the best fitting of the experimental scattering intensity [25, 48-50]. A MD simulation using a SAXS-based pseudo energy function requires the analytic derivative of the potential with respect to Cartesian coordinates. Moreover, having an analytic derivative will also greatly facilitate the mathematical optimization process using a pseudo energy function. However, for many SAXS modeling programs such as Crysol and Fast-SAXS-pro, the analytical derivative of the hydration term is not available, making it difficult to implement a derivative-based structure optimization. Consequently, Björling et al. [48] and Kimanius et al. [49] simply ignored the effect of hydration shell in their SAXS-guided MD simulations, making it difficult to directly compare their results with experimental data. As an alternative solution, Zheng et al. [25] computed the numerical derivative of a SAXS-fitting score and performed structure optimization using the Newton-Raphson algorithm. Extending Zheng's work for MD simulation is challenging because of the lack of analytical derivative, and the hydration modeling in the approach is empirical with room for further improvements. Very recently, Chen et al. [50] computed the analytical derivative of a

SAXS-biased potential for MD simulations. Chen's approach can be time-consuming due to the required explicit solvent MD simulations, and it is mainly applied for near-WAXS high resolution data. Therefore, it is still of great interest to develop a fast yet accurate expression for hydration scattering with the analytical derivative, for efficient structure optimization fitting experimental SAXS profiles.

In this work, we intend to propose an accurate approach to calculate hydration layer scattering, while eliminating expensive computations using an explicit water model. The analytical derivative with respect to Cartesian coordinates is available in our SAXS computing expression, which is particularly advantageous for gradient-based optimization and SAXS-guided MD simulations. Inspired by the FoXS program, [21, 22] the hydration scattering intensity is evaluated using only the coordinates of protein atoms with some modified form factors, and the implicit hydration contribution is proportional to the solvent accessible surface area (SASA) of a protein atom. The scattering intensities computed using all-atom models together with explicit hydration water molecules are used as the fitting target to parameterize these modified form factors incorporating hydration. Two separate sets of modified form factors are developed, for atomistic and C α -based coarse-grained (CG) protein models, respectively. The CG approach is based on our recent work on residue-level form factors [92], and using CG structure models can further improve the computational efficiency. After comparison to the results using explicit water, it is shown that our implicit hydration modeling is of high accuracy. Furthermore, by combining our modeling approach with a conjugate gradient method, the SAXS-based structure optimization produces good results for several selected proteins.

5.1.1 Computation of target scattering intensities

The target scattering intensities are computed using all-atom protein structures with explicit hydration water molecules by equation (2.7), which is rewritten here for clarity.

$$I(q) = \left\langle \left| A_p(\mathbf{q}) - A_e(\mathbf{q}) + A_s(\mathbf{q}) \right|^2 \right\rangle_{\Omega} \quad (5.1)$$

In the above equation, $A_p(\mathbf{q})$ is the scattering amplitude of protein atoms *in vacuo*. $A_e(\mathbf{q})$ is the scattering amplitude of dummy atoms representing the excluded volume effect of protein atoms. The third term represents the contribution of excess hydration density at the protein surface compared with bulk solvent, and $A_s(\mathbf{q})$ is the scattering amplitude of the water molecules in the hydration shell. $\langle \dots \rangle_{\Omega}$ represents the orientational average over different scattering vector $\mathbf{q}$ with the same scalar value. We choose 1000 directions uniformly distributed on the $\mathbf{q}$ space sphere for each scalar value to ensure a high accuracy computation. The directions are chosen based the spiral method following Park et al. [32].

The excluded volume effect is usually considered by modifying the protein atomic form factors using the Gaussian sphere approximation following Fraser et al. [36], as shown in equation (2.2). Thus the first two terms can be written as

$$A_p(q) - A_e(q) = \sum_j f_j^{excl}(q) \exp(-i\mathbf{q}\mathbf{r}_j) \quad (5.2)$$

where j loops over all the protein atoms.

Additionally, the hydration effect can be taken into consideration by using explicit hydration water molecules. Similar to equation (5.2), the third term can be computed using the following formula.

$$A_s(q) = w \cdot \sum_k f_k(q) \exp(-i\mathbf{q}\mathbf{r}_k) \quad (5.3)$$

where k loops overall all the atoms in hydration water molecules and w is the weighting factor determining the contribution of the hydration layer. In order to obtain the equilibrium structure of water around the protein, a short molecular dynamics (MD) simulation of 100 ps was performed for each protein structure in our dataset. The simulations were performed using Gromacs 4.5.5 [55] with CHARMM27 all-atom force field [56] and the SPC/E water model [65] was used. The hydration water molecules are then selected to be those within 4.5 Å from the protein atoms. This approximately corresponds to a hydration shell with thickness of 3 Å, as the van der Waals radii of protein atoms are around 1.5 Å. The thickness of hydration shell is consistent with most methods in literature. The weight of hydration water molecules is selected to be 3%, following Yang et al. [29] and Zheng et al. [25]. Some other weight values were also tested and it is found that the optimization results of the implicit hydration term were approximately proportional to the weight of hydration layer.

5.1.2 Evaluation of scattering intensity using the implicit hydration term

The above method to compute the scattering intensity is not suitable for a gradient-based structure optimization or a SAXS-guided MD simulation. This is because the positions of hydration water molecules need to be re-evaluated every time when the structure is changed and the analytical derivative with respect to Cartesian coordinates is not available. In the following, we propose to

evaluate the scattering intensity from the protein structure only so that the derivative with respect to protein coordinates can be easily obtained.

Here the scattering intensity is still computed using equation (5.1). However, the third term $A_s(\mathbf{q})$ is now computed using the coordinates of protein atoms instead of hydration water molecules, as shown in the following

$$A_s(\mathbf{q}) = \sum_j s_j f^s(q) \exp(-i\mathbf{q}\mathbf{r}_j) \quad (5.4)$$

where s_j is the solvent accessible surface area (SASA) of protein atom j and $f^s(q)$ is the correction to the form factor considering the hydration effect. j loops over all the protein atoms. This approach corresponds to that the form factors of protein atoms are modified to be in the following form.

$$f_j^{new}(q) = f_j^{excl}(q) + s_j f^s(q) \quad (5.5)$$

where $f_j^{excl}(q)$ is the protein atomic form factor subtracted by the form factor of dummy atom representing the excluded solvent, as shown in equation (2.2). In equation (5.4) and (5.5), the contribution of hydration water is assumed to be proportional to the SASA only, irrelevant to the atom type. This assumption is generally true as we consider the scattering intensity of SAXS region only. A similar expression of the hydration term can be found in the FoXS program [21, 22].

5.1.3 Optimization of the hydration term for atomistic models

The implicit hydration term $f^s(q)$ is optimized so that the scattering intensities computed using our implicit hydration modelling method best reproduce those

from explicit hydration calculations. The following residual is defined to measure the difference between the two methods for all the protein structures in our dataset.

$$W(q) = \sum_i^N \left(\frac{I_i(q) - I_i^0(q)}{I_i^0(q)} \right)^2 \quad (5.6)$$

Here, $I_i^0(q)$ is the target scattering intensity computed using explicit hydration water molecules and $I_i(q)$ is the scattering intensity computed using modified atomic form factors shown in equation (5.5), where $f^s(q)$ is used to represent the contributions of the hydration shell. Index i loops over all the protein structures in our dataset. In order to obtain the optimal $f^s(q)$ values, we implement the Newton's method to minimize the residual, as the secondary order of derivative of $W(q)$ to $f^s(q)$ can be easily computed. Starting from $f^s(q) = 0$ for all q values, $f^s(q)$ is updated using the following equation.

$$\Delta f^s(q) = - \left(\frac{\partial W(q)}{\partial f^s(q)} \right) / \left(\frac{\partial^2 W(q)}{\partial f^s(q)^2} \right) \cdot \lambda \quad (5.7)$$

Here, λ is the step size of updating, which is chosen to be 0.5 here. The optimization converges when the relative changes of all $f^s(q)$ values are smaller than 0.001. The first and second order of derivative of $W(q)$ with respect to $f^s(q)$ can be written in analytical form as derived in the following. For simplicity, we just removed the variable q in parenthesis in the following equations.

$$\begin{aligned}
\frac{\partial W}{\partial f^s} &= \sum_i^N \frac{2}{(I_i^0)^2} (I_i - I_i^0) \frac{\partial I_i}{\partial f^s} \\
\frac{\partial^2 W}{\partial (f^s)^2} &= \sum_i^N \frac{2}{(I_i^0)^2} \left[\left(\frac{\partial I_i}{\partial f^s} \right)^2 + (I_i - I_i^0) \frac{\partial^2 I_i}{\partial (f^s)^2} \right] \\
\frac{\partial I_i}{\partial f^s} &= \left\langle 2 \sum_j (f_j^{excl} + s_j f^s) \cos(\mathbf{q}\mathbf{r}_j) \sum_j s_j \cos(\mathbf{q}\mathbf{r}_j) \right\rangle_{\Omega} + \\
&\quad \left\langle 2 \sum_j (f_j^{excl} + s_j f^s) \sin(\mathbf{q}\mathbf{r}_j) \sum_j s_j \sin(\mathbf{q}\mathbf{r}_j) \right\rangle_{\Omega} \\
\frac{\partial^2 I_i}{\partial (f^s)^2} &= \left\langle 2 \left[\sum_j s_j \cos(\mathbf{q}\mathbf{r}_j) \right]^2 + 2 \left[\sum_j s_j \sin(\mathbf{q}\mathbf{r}_j) \right]^2 \right\rangle_{\Omega}
\end{aligned} \tag{5.8}$$

5.1.4 Optimization of the hydration term for coarse-grained models

Apart from all-atom structures, the same fitting approach is applied to the widely used C α -based coarse-graining scheme. The basic procedure is similar to the above all-atom case. However, during the fitting process, it is desirable to separate the error of scattering intensity computation caused by the coarse-grained form factors and the approximation of using $f^s(q)$ to represent hydration shell contribution. So for the first two terms in equation (5.1), the values computed from all atom structure are still used, and the third term $A_s(\mathbf{q})$ is evaluated from CG protein coordinates as follows.

$$A_s(\mathbf{q}) = \sum_J s_J f^s(q) \exp(-i\mathbf{q}\mathbf{R}_J) \tag{5.9}$$

Here, J loops over all the CG beads (C α atoms in this work) in the protein structure. s_J is the SASA of J^{th} CG bead. For the C α CG scheme used here, s_J

can be directly evaluated using the empirical parameters developed by POPS [93].

It should be noted that the spherical average method to compute scattering intensity is implemented here, instead of the commonly used Debye formula. The main reason is that the scattering amplitude can be easily separated to different parts. Thus, it is possible to combine the $A_p(\mathbf{q})-A_e(\mathbf{q})$ term computed from all atom structures with the $A_s(\mathbf{q})$ term computed from the CG counterparts during our optimization of CG $f(q)$. This separation guarantees that the resulting $f(q)$ is not affected by the errors introduced in the CG form factors development in our previous work [92].

5.2 Structure optimization using the optimized hydration term

The advantage of the implicit hydration term introduced above is that the contribution of the excess hydration shell can be estimated using the protein atomic coordinates directly. Thus, we can easily obtain the analytical derivative of the scattering intensity to the protein coordinates. This allows us to use more sophisticated algorithms to optimize the protein structure for better fitting of the given scattering intensity curve. We have presented using the conjugate gradient method, which is a first-order derivative based optimization method, to perform structure optimization in this part. To be convenient, we select to use the one-bead C α coarse-grained representation instead of all-atom structure.

5.2.1 Pseudo energy function as optimization target

A pseudo energy function containing both a structure-constraint term and a SAXS-fitting term is defined in order to perform the structure optimization. In this work, we focus on the C α -based structure optimization to illustrate the usage of the analytical derivative. The form of the pseudo energy function is shown in the following equation.

$$W = (1 - \lambda)cW_s + \lambda W_q \quad (5.10)$$

Here, W_s is the structure-constraint term and W_q is the SAXS-fitting term. λ is the weight of the SAXS-fitting term in the energy function, which increases generally during the optimization process. As the two energy terms are computed using very different formulas, a scaling parameter c is added to ensure that they are approximately in the same order of magnitude. The standard G δ -like model [94] is used as the structure constrain term W_s , which contains the following several parts.

$$\begin{aligned} W_s &= E_{bonded} + E_{contact} + E_{rep} \\ &= \sum_{bonds} k_b (r - r^0)^2 + \sum_{angles} k_\theta (\theta - \theta^0)^2 + \sum_{dihedrals} k_\phi F(\phi) \\ &\quad + \sum_{contacts} k_c \left[5 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 6 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{10} \right] + \sum_{non-contacts} k_{nc} \left(\frac{\sigma_0}{r} \right)^{12} \end{aligned} \quad (5.11)$$

E_{bonded} is the energy of bonded terms, including all bonds, angles and dihedrals. For bonds and angles, harmonic potential functions are used to restrain the values in the vicinity of the initial structure (r^0 and θ^0). The cosine functions are used for dihedrals, where $F(\phi) = [1 - \cos(\phi - \phi^0)] + \frac{1}{2} [1 - \cos(3(\phi - \phi^0))]$

and the superscript 0 indicate that the dihedral value comes from the initial structure. $E_{contact}$ is the energy of native contact pairs presented in the initial structure, which can be evaluated using the shadow map algorithm [95]. The native interaction is modeled by a 12-10 Lennard-Jones potential here instead of the commonly used 12-6 potential in atomistic MD simulations following the Gō-like model [94]. σ_{ij} is the distance between two native-contact beads presented in the initial structure. At last, the pure repulsive potential E_{rep} is used to prevent that non-contact CG bead pairs become too close. σ_0 is chosen to be 4 Å following Clementi et al. [94]. The force constant of the interactions present above are set to be $k_b = 100 \text{ kJ/mol Å}^{-2}$, $k_\theta = 20 \text{ kJ/mol rad}^{-2}$, $k_\phi = 1 \text{ kJ/mol}$, and $k_c = k_{nc} = 1 \text{ kJ/mol}$. The structural information contained in the one-dimensional SAXS curve is usually not enough to completely determine the protein structure. In this case, using a structure-constraint potential can efficiently eliminate many unreasonable structures. Nevertheless, the Gō-like potential used here would bias the optimized structure towards the initial structure. Thus, the optimization result will be better when the initial structure and the optimization target have similar local structures, such as secondary structure distributions. Additionally, the conjugate gradient algorithm used in this work is a local optimization method, and it is possible that the optimization is trapped in a local minimum. Nevertheless, reasonable structure optimization results are obtained for the test cases, as shown in section 3.4. To further improve the accuracy of the structure optimization process, more sophisticated structure-restraint potential and a global optimization algorithm based on MD simulation or simulated annealing should be applied.

The SAXS-fitting score W_q measures how well the scattering intensity of current structure fits the target scattering intensity. W_q is simply defined following the widely used χ^2 measurement, which has the following format.

$$W_q = \frac{1}{N_q} \sum_q \left[\frac{f \cdot I(q) - I_0(q)}{\sigma(q)} \right]^2 \quad (5.12)$$

where N_q is the number of points on the scattering intensity curve, $I(q)$ is the scattering intensity computed from the current structure and $I_0(q)$ is the target scattering intensity. $\sigma(q)$ is the standard error of the target scattering intensity. In our test cases of structure optimization, the target scattering intensity is the theoretical scattering intensity computed from a specific target structure, and no experimental standard error is available. Thus, an empirical standard error value is used here following Stovgaard et al. [42], as shown in equation (2.12).

Using the $f^s(q)$ obtained in our optimization method, the scattering intensity of a coarse-grained protein structure can be computed using the Debye formula as follows.

$$I(q) = \sum_i \sum_j \left[f_i^{CG}(q) + s_i f^s(q) \right] \left[f_j^{CG}(q) + s_j f^s(q) \right] \frac{\sin(qr_{ij})}{qr_{ij}} \quad (5.13)$$

Here, $f^{CG}(q)$ is the CG form factors derived in our previous work [92] and s is the solvent accessible surface area (SASA) of the CG beads. Using the standard approach in literature [93, 96], the SASA value s is computed from the protein coordinates using

$$s_i = A_i \cdot \prod_{j=1}^N \left[1 - \frac{p_i p_j b_{ij}(r_{ij})}{A_i} \right] \quad (5.14)$$

A_i is the surface area of bead i , which can be computed by $A_i = 4\pi(R_i + R_{solv})^2$. R_i is the van der Waals radius of CG bead i . R_{solv} is the radius of solvent molecules, chosen to be 1.4 Å. The product term means the reduction of the SASA value of bead i caused by the overlapping with other CG beads. p_i is a SASA parameter specified by bead type and p_{ij} is an overlapping parameter depending on the bond connection between CG bead i and j . $b_{ij}(r_{ij})$ describes the overlapping area of two spherical beads i and j , with the radii R_i , R_j and distance r_{ij} . When $r_{ij} > R_i + R_j + 2R_{solv}$, the two atoms do not have overlap, thus $b_{ij}(r_{ij}) = 0$, else $b_{ij}(r_{ij}) = \pi(R_i + R_{solv})(R_i + R_j + 2R_{solv} - r_{ij}) \cdot \left[1 + (R_j - R_i)/r_{ij}\right]$. The values of all parameters in equation (15) are taken from the POPS program [93].

5.2.2 Iterative optimization based on analytical derivatives

From equation (5.13) and (5.14) shown above, all the terms included in the computation of scattering intensity are analytically represented by the protein coordinates. Thus, the analytical derivative of the scattering intensity with respect to the protein coordinates can be obtained. Furthermore, the analytical derivative of the pseudo energy function defined in equation (5.11) can also be obtained. The detailed expressions of the analytical derivatives are presented in the following equations.

$$\begin{aligned}
\frac{\partial W}{\partial x_i} &= (1-\lambda)c \frac{\partial W_s}{\partial x_i} + \lambda \frac{\partial W_q}{\partial x_i} \\
\frac{\partial W_s}{\partial x_i} &= \sum_{bonds} 2k_b (r_{ij} - r_{ij}^0) \frac{x_i - x_j}{r_{ij}} + \sum_{angles} 2k_\theta (\theta - \theta^0) \frac{\partial \theta}{\partial x_i} - \sum_{dihedrals} k_\phi \frac{\partial F(\phi)}{\partial x_i} \\
&\quad \left[\sum_{contacts} 60k_c \left(\frac{\sigma_{ij}^{10}}{r_{ij}^{11}} - \frac{\sigma_{ij}^{12}}{r_{ij}^{13}} \right) + \sum_{non-contact} 12k_{nc} \frac{\sigma_0^{12}}{r_{ij}^{13}} \right] \frac{x_i - x_j}{r_{ij}} \\
\frac{\partial F(\phi)}{\partial x_i} &= \left[\sin(\phi - \phi^0) + \frac{3}{2} \sin(3\phi - 3\phi^0) \right] \frac{\partial \phi}{\partial x_i} \\
\frac{\partial W_q}{\partial x_i} &= \frac{1}{N_q} \sum_q \frac{2}{\sigma^2(q)} [I(q) - I_0(q)] \frac{\partial I(q)}{\partial x_i} \\
\frac{\partial I(q)}{\partial x_i} &= \sum_i \sum_j \left\{ \left[\frac{\partial s_i}{\partial x_i} f_j(q) + \frac{\partial s_j}{\partial x_i} f_i(q) \right] f^s(q) \frac{\sin(qr_{ij})}{qr_{ij}} + \right. \\
&\quad \left. f_i(q) f_j(q) \left[\frac{\cos(qr_{ij})}{qr_{ij}} - \frac{\sin(qr_{ij})}{qr_{ij}^2} \right] \frac{x_i - x_j}{r_{ij}} \right\} \\
\frac{\partial s_i}{\partial x_i} &= \sum_{j=1}^N s_i \left(\frac{A_i}{p_i p_{ij}} - b_{ij}(r_{ij}) \right)^{-1} \frac{db_{ij}(r_{ij})}{dr_{ij}} \frac{x_j - x_i}{r_{ij}} \\
\frac{\partial s_i}{\partial x_i} &= s_i \left(\frac{A_i}{p_i p_{ij}} - b_{ij}(r_{ij}) \right)^{-1} \frac{db_{ij}(r_{ij})}{dr_{ij}} \frac{x_j - x_i}{r_{ij}} \\
\frac{db_{ij}(r_{ij})}{dr_{ij}} &= -\pi (R_i + R_{solv}) \left[1 + \frac{(R_i + R_j + 2R_{solv})(R_j - R_i)}{r_{ij}^2} \right]
\end{aligned} \tag{5.15}$$

Similar equations can be written for the derivatives of coordinates y_i and z_i . This is a great advantage as some advanced optimization algorithms can be implemented in our structure optimization process. Here, the standard conjugate gradient algorithm following Numerical Recipe [97] is implemented. The procedure of the iterative optimization process is described in as follows.

1. The initial value of λ is set to be 0.5.

2. The current pseudo energy function is optimized by the conjugate gradient algorithm. The convergence criterion is set to be 10^{-5} .

3. Set $\lambda = 1 - 0.5^n$ and repeat step 2, where n is the iteration cycle.

4. The iteration ends if the number of steps in each iteration cycle is fewer than 10 for three consecutive iteration cycles or the maximum number of 20 iterations is reached.

Step 4 in the above procedure is devised to achieve the optimal fitting quality and the detailed discussion can be found in Section 5.6.

5.3 Resulting implicit hydration term

In the structure optimization process, it is desired to have a simple expression of scattering intensity, whose derivatives with respect to the structure coordinates can be obtained analytically. Although many methods in literatures can compute scattering intensity in reasonable accuracy, most of the methods, especially those using explicit positions of hydration water molecules, do not have analytical derivatives. To efficiently compute the scattering of hydration layer, we propose to directly modify the form factors of protein atoms by incorporating an implicit hydration term multiplied by the solvent accessible surface area (SASA) of the protein atoms (or residues at the coarse-grained level). The implicit hydration term is fitted so that the computed scattering intensities of proteins in our dataset can best reproduce those computed with explicit hydration water molecules. In this way, the computation of scattering intensity is related to the protein coordinates only, and still preserves reasonable

accuracy. The SAXS-based structure optimization can thus be more convenient than many methods available in literature because of the analytical derivatives.

Figure 5.1 Comparison of the implicit hydration term $f^s(q)$ optimized using atomistic and C α -based coarse-grained protein structures. The result of all-atom representation is shown in red curve, while the C α based coarse-grained scheme is shown in blue curve.

The implicit hydration term is optimized for all-atom as well as CG protein structures using the procedure described in the Methods part. Figure 5.1 shows the resulting implicit hydration term $f^s(q)$ computed using both structure representations. It is seen that the two curves show similar trends, although there are some deviations at the relatively high q region. The values of $f^s(q)$ equals approximately to 0.03 at $q = 0$, which is consistent to the 3% excess electron density of the 3Å hydration shell used in the fitting ($0.334 \times 3 \times 0.03$), with the bulk electron density 0.334 e/Å^3 . The $f^s(q)$ curve shows a local minimum at about $q = 0.19 \text{ Å}^{-1}$, where its value is approximately zero. This

means the contribution of hydration layer is approximately 0 at this q value. After that, the value of $f^s(q)$ increases until about $q = 0.25 \text{ \AA}^{-1}$ before it starts decreasing. The high q part of the CG $f^s(q)$ curve is flatter than the atomistic counterpart and the corresponding points are closer to zero, which may be due to the factor that coarse-graining eliminates some high-resolution structural features. The overall similarity of curve shape between the resulting implicit hydration terms for different resolutions of structure representation also indicates that the hydration scattering expression based on SASA works for protein models at different coarse-graining levels.

It is also interesting to compare our result with the similar term used by FoXS [21, 22]. The hydration scattering in this work is proportional to the SASA values, instead of the SASA fractions used in FoXS. Using SASA fractions indicates a fixed number of neighboring water molecules around a fully exposed solute atom or a CG site, which is a reasonable approximation for atomistic form factors. However, as different CG beads can have significantly different sizes, the expression based on SASA values are adopted in this work for both atomistic and CG levels of computations. Furthermore, the $f^s(q)$ term in our approach is a tabulated function that allows the full flexibility in the form factor optimization, while FoXS assumes that the effective scatters acting on protein surface atoms still maintain the water form factor values.

5.4 Accuracy of SAXS profile evaluation

In order to estimate the accuracy of the implicit hydration terms obtained above, two sets of target scattering profiles are used. The first set contains the theoretical scattering profiles computed from all-atom protein structures in our

constructed protein dataset with explicit hydration water molecules, and the second set consists of some selected experimental scattering profiles. The first set of theoretical target profiles are used to evaluate the accuracy of a scattering modeling approach, while additional validation is performed using the second set of experimental data. The scattering intensities of the two sets of proteins were evaluated using our optimized implicit hydration terms (for both atomistic and CG approaches). Additionally, theoretical profiles using Crysol (atomistic approach) and Fast-SAXS-pro (CG approach) were also computed using the first set of proteins for comparison. The differences between scattering profiles are measured by χ^2 or R values, as defined in equation (2.10) and (2.11), depending on the target scattering profiles. Smaller χ^2 or R values indicate better fitting results. The protein scattering amplitudes (including the excluded solvent contribution) for atomistic and protein computations come from the conventional atomistic approach [36, 98] and our previously developed protein CG form factors [92], respectively.

For the 880 high resolution protein structures in our dataset, the distributions of R values are shown in Figure 5.2, using the theoretical profiles from atomistic protein models and explicit-solvent as the target.

Figure 5.2 The R value distributions of scattering intensities computed from our dataset containing 880 protein structures. The R values are computed using the scattering intensities evaluated by atomistic protein structures plus explicit hydration layers as targets. (a) Atomistic structure representations are used to compute the scattering intensities. The red and blue curves show the distributions of our implicit hydration treatment and Crysol respectively. (b) $C\alpha$ -based CG structure representations are used to compute the scattering intensities. The red and blue curves show the distributions of our implicit hydration treatment and Fast-SAXS-pro respectively.

Figure 5.2(a) illustrates the results computed from atomistic structure representations, where the R value distributions of our implicit hydration treatment and Crysol are shown in red and blue curves respectively. The excess electron density in Crysol is also chosen to be 3% of the bulk to be consistent with the theoretical calculation for the fitting target. It is shown in the figure that our implicit hydration treatment is slightly closer to the explicit-solvent

approach than Crysol. The major peak in the R value distribution appears in the range of 0 ~ 0.2 from our method, while that appears in the range of 0.2 ~ 0.4 in the Crysol result. Note that the explicit- solvent approach used as the target in this work is not necessarily superior to Crysol, as both methods have a number of approximations. Nonetheless, the small R values for our implicit-solvent approach and Crysol manifest that both methods generate results close to the target, which is a validation of our implicit-solvent approach.

The results of C α -based CG approaches using our implicit hydration treatment and Fast-SAXS-pro are plotted in Figure 5.2(b). Generally, the R values are larger than those from the atomistic approaches shown above. Although not as accurate as atomistic methods such as Crysol, our CG approach using the implicit hydration treatment outperforms Fast-SAXS-pro. A large portion of R values (82.6%) from our CG approach are smaller than 1.0 in our R value distribution, suggesting an acceptable precision. Finally, results by FoXS were also calculated (data not shown), using the atomistic and CG options of the software. While FoXS atomistic result is similar to that by Crysol, the R value distribution from the FoXS CG approach is obviously worse than that by Fast-SAXS-pro, with the major peak in the range of 2.0 ~ 3.0. The unsatisfactory FoXS CG result may be due to using the fraction of SASA as discussed above. Our implicit hydration CG treatment produces significantly more accurate results compared with existing methods with the same CG mapping, because the implicit hydration term used in our treatment is obtained from fitting the target.

Figure 5.3 The comparison of SAXS profiles computed using explicit hydration water molecules and our implicit hydration term, as well as Crysol, for four selected proteins with different sizes. The red spheres are the all-atom SAXS profiles computed using explicit hydration water molecules. The green lines represent the all-atom SAXS profiles evaluated with the atomic form factors modified by our implicit hydration term and the blue curves illustrate the results computed from Crysol. The four example proteins shown in the figures are: (a) 4CO8 (98 residues), (b) 4JL5 (203 residues), (c) 4UYB (400 residues) and (d) 1N62 (804 residues).

Figure 5.4 The comparison of SAXS profiles computed using explicit hydration water molecules and our implicit hydration term, as well as Fast-SAXS-pro for four selected proteins with different sizes. The red spheres are the all-atom SAXS profiles computed using explicit hydration water molecules. The green lines represent the coarse-grained SAXS profiles evaluated with the residue-based coarse-grained form factors modified by our implicit hydration term and the blue lines illustrate the results of Fast-SAXS-pro. The four example proteins shown in the figures are: (a) 4CO8 (98 residues), (b) 4JL5 (203 residues), (c) 4UYB (400 residues) and (d) 1N62 (804 residues).

To further visualize the performance of our implicit hydration treatment, protein scattering intensities computed using our approach and explicit hydration water molecules are plotted. Figure 5.3 shows the results of four example proteins, and the Crysol results are also plotted for comparison. The

all-atom protein structures are used to evaluate the protein scattering amplitudes. The four proteins are chosen to have different numbers of residues. The PDB codes and sizes are 4CO8 (98 residues) [86], 4JL5 (203 residues) [87], 4UYB (400 residues) (To be published) and 1N62 (804 residues) [88]. From all the four plots, it is seen that all the scattering intensities overlap with the target profiles very well. For the first two proteins, there are small deviations at the q range of around $0.25 \text{ \AA}^{-1}$, as reflected by the slightly larger R values. It is also found that the R values from Crysol are small for the four proteins, indicating that all of the four atomistic approaches shown in Figure 5.3 generate very close scattering profiles. Similarly, the scattering profiles of our CG approach for the same set of proteins are plotted in Figure 5.4 and compared with the all-atom explicit solvent targets and the results of Fast-SAXS-pro. As shown in Figure 5.4(a), our implicit hydration approach with a $C\alpha$ -based protein model does not work perfectly for this small protein, with an R value of 1.126, and the derivation from the target profile starts in the $0.25 - 0.3 \text{ \AA}^{-1}$ region. Nevertheless, the comparison with Fast-SAXS-pro shows significant improvement using our implicit hydration CG approach for the protein in Figure 5.4(a) and all other three proteins. The implicit hydration CG method works better for the two larger proteins in Figure 5.4(c) and (d) with 400 and 804 residues respective. This phenomenon is further supported by the analysis on all proteins in our protein database, which shows that among the 152 proteins with R values greater than 1.0, only three proteins have more than 400 residues. Based on the above results, it is suggested that the CG version of our approach should be only applied for a low- q region (e.g., $q < 0.3 \text{ \AA}^{-1}$), especially for small proteins. Apart from applying the all-atom approach, higher resolution models such as a two-bead-per-residue model need to be developed for modeling SAXS data of higher q region.

Table 5.1 Fitting of experimental SAXS profiles using our implicit hydration treatment (with both an atomistic and a $C\alpha$ based coarse-grained structure representations) and Crysol for six selected proteins. The experimental SAXS profiles are obtained from SASBDB, with the entry of each protein listed in the table. The weights of hydration contribution and protein excluded volume are tuned to best reproduce experimental data.

		w_{wat} (%)	w_{excl}	χ^2
Lysozyme (SASDAC2)	AA	1.392	1.011	1.434
	$C\alpha$	0.775	1.052	1.320
	Crysol	1.5	1.025	1.404
Ubiquitin (SASDAQ2)	AA	1.780	1.015	1.416
	$C\alpha$	1.541	1.091	2.561
	Crysol	3.0	1.074	1.766
Myoglobin (SASDAK2)	AA	2.201	1.023	0.948
	$C\alpha$	0.848	1.088	1.193
	Crysol	2.1	1.000	0.984
Rnase (SASDAN2)	AA	4.014	1.014	3.017
	$C\alpha$	2.743	1.049	2.248
	Crysol	5.4	0.914	1.582
Cyt-C (SASDAB2)	AA	0.045	0.982	1.640
	$C\alpha$	0.000	0.948	1.500
	Crysol	1.5	0.975	1.040
Chymotrypsinogen A (SASDAA8)	AA	2.934	0.986	2.382
	$C\alpha$	2.130	0.976	2.472
	Crysol	3.9	0.951	2.726

Additionally, six proteins with experimental SAXS profiles in SASBDB [15], are selected to test the accuracy of our fitted implicit hydration terms. The six chosen proteins are Lysozyme (PDB ID: 1LYZ) [59], Ubiquitin (PDB ID: 1UBQ) [89], Myoglobin (PDB ID: 1WLA) [60], RNase (PDB ID: 1C0B) [61], Cytochrome C (PDB ID: 1HRC) [90] and Chymotrypsinogen A (PDB ID: 2CGA) [91]. The entries of the experimental data for all the proteins are included in Table 5.1. To compare with experimental SAXS profiles, the weights of hydration layer and protein excluded volume are optimized in both our method and Crysol. The final χ^2 values, as well as the optimal weight values, are presented in Table 5.1. From the Table, it is found the χ^2 values of different methods are similar for the six proteins. Crysol performs slightly better for Lysozyme, RNase and Cytochrome C, while our implicit atomistic or CG method produces smaller χ^2 values for other proteins. In a few cases, the χ^2 values of the C α -based CG approach are even smaller than the atomistic counterparts when both using our implicit hydration treatment. This indicates the possible over-fitting when comparing with real experimental SAXS profiles. For example, the weight of hydration term is optimized to be 0 for protein Cytochrome C using the CG approach, after applying the search boundary 0-20% similar to 0-0.075 e/A³ used in Crysol.

Taken the above two parts of results targeting at theoretical and experimental profiles together, our method to modify the protein form factors with an implicit hydration term can obtain accurate SAXS profiles on the basis of the protein coordinates only. The accuracy of our atomistic method is comparable to the uniform hydration layer treatment used in Crysol, while the accuracy our CG approach is only slightly lower than that of Crysol and is significantly higher than that of some previously published CG methods.

5.5 Robustness of the fitting process

It is demonstrated here that our fitting result obtained above is convergent and robust. The all-atom structure is used to perform the test. The result remains nearly unchanged when different protein datasets are used to perform the optimization process. It is further shown that the obtained implicit hydration term scales almost linearly with the hydration weight value used to compute the fitting target scattering intensities at relatively small values of hydration weight.

First, to validate that the R value distributions obtained above are converged and not due to over-fitting, additional optimizations of the implicit hydration term $f^s(q)$ were performed using different numbers of protein structures, ranging from 100 to 800 and randomly selected from the whole dataset. The obtained $f^s(q)$ curves are plotted and compared in Figure 5.5(a). Four curves, with $N = 200, 400, 600$ and 880 (used as target), are selected to show the excellent overlapping. In order to give a quantitative measurement of the overlap between two $f^s(q)$ results obtained using different number of protein structures, the relative difference is defined by the formula $d = \frac{1}{N_q} \sum_q \left(\frac{f^s(q) - f_0^s(q)}{f_0^s(q)} \right)^2$. In

the equation, N_q is the number of data points in the curve, $f_0^s(q)$ is the implicit hydration term computed from the whole dataset and the denominator is the average value of $f_0^s(q)$. Here, $f_0^s(q)$ is not used as the denominator directly, because some values are very close to zero and will lead to numerical instability. The relative difference values for all the $f^s(q)$ compared with that computed using the whole dataset are plotted in Figure 5.5(b). From the figure, it is seen that the difference values are very small. Even when as few as 100 protein

Figure 5.5 (a) the optimized implicit hydration term using different numbers of protein structures in our dataset. The result of using the whole dataset (880 structures) is shown in red curve, as the comparison target for computing relative differences. The results of using 200, 400, 600 protein structures to perform optimization are shown in blue, magenta and cyan respectively. (b) The relative differences of the resulting implicit hydration terms computed using different numbers of protein structures from 100 to 800. The comparison target is the red curve shown in (a). The formula to compute the relative difference can be found in the main text.

Figure 5.6 Comparison of implicit hydration terms optimized when different hydration weight is used to calculate the target scattering profiles. For the comparison purpose, the resulting implicit hydration terms are divided by their corresponding hydration weight values respectively.

structures are used to perform the optimization, the difference value, which means the relative error of the curve, is less than 1.2%. It is also found for 400 or more protein structures, the relative difference values become nearly unchanged. For this protein number range, the relative difference mainly represents the numerical errors during the fitting process. Thus it is concluded our method to optimize the implicit hydration term tends to converge at about 400 protein structures. The implicit hydration term obtained above using the whole dataset is a numerically reliable result.

Next, different values of hydration weight to calculate the target scattering profiles are tested to check the dependence of our optimization result to the hydration weight. The value 3% used in our computation is an empirical chosen value following research many approaches in literatures [25, 29], assuming that the scattering amplitude of the hydration layer is 103% of that of the same

water layer filled by bulk water. However, different values of the hydration layer weight have also been reported [20, 31]. Here, tests were performed to find out how our optimized implicit hydration term varies according to the value of the hydration layer weight used. Weight values of 3%, 5%, 7% and 10% were selected. In order to compare the hydration terms of different weights, the resulting implicit hydration terms are divided by their corresponding weight values and are plotted in Figure 5.6. From the figure, it is seen that different implicit hydration terms are very close after the scaling. The relative difference values of the later three hydration weight values, compared with that obtained by hydration weight 3%, are 0.8%, 3.1% and 9.8% respectively, according to the formula to measure differences mentioned above. These small difference values mean the hydration term obtained using our optimization method is almost linearly related to the weight of the hydration layer used to compute the target scattering intensity, in case that the weight value is relatively small. For values around or below 5%, which is the case for most proteins, the differences are indeed negligible. Thus, the predefined weight (3%) for the explicit-solvent fitting target does not affect the generality of our optimized implicit hydration scattering term, because results from a different weight can be easily obtained by simple scaling. In SAXS profile modeling, the hydration scattering contribution is usually scaled by a constant for better fitting to experimental data. Therefore, our assumption of the 3% weight has in principle no effect on experimental data fitting.

In conclusion, our optimization method and the resulting $f^s(q)$ term obtained is a converged and numerically stable result. The resulting hydration term also scales with the weight of hydration water molecules almost linearly. Thus it can be easily extended to more generalized cases. The conclusion is based on the

atomistic approach, and tests for the CG scheme resulted in the same conclusion (data not shown).

5.6 Result of structure optimization

As discussed in the method part, the advantage of incorporating the contribution of hydration layer into the form factors of protein atoms (or CG sites) is the availability to obtain the analytical derivatives of the scattering intensity with respect to the protein coordinates. Here, this advantage is demonstrated by performing a structure optimization given a target scattering profile and an initial structure. The CG one-bead per residue C α scheme is used here for its convenience. The pseudo energy function to be minimized and the procedure to perform optimization are shown in the method part. As the value of the hydration layer weight needs to be fixed during the optimization, the implicit hydration term fitted from 3% hydration weight is implemented following the empirical value in literature. From the discussion in Section 3.2, the CG approach of our implicit hydration treatment tends to be accurate for q range of $0 \sim 0.3 \text{ \AA}^{-1}$. So our structure optimization is also performed based on this q range.

As the test cases, structure optimizations toward known target structures were performed to validate our method. Four proteins that have different conformations in the Protein Data Bank (PDB) were selected as benchmark systems. They are adenylate kinase (PDB ID: 1AKE, 4AKE) [99, 100], LAO binding protein (PDB ID: 1LST, 2LAO) [101], maltodextrin-binding protein (PDB ID: 1OMP, 1ANF) [60, 102] and elongation factor 2 (PDB ID: 1N0V, 1N0U) [103]. For each protein, the first structure was selected as the target

structure and the scattering intensity computed from this structure was set to be the target scattering intensity. Structure optimizations were then performed using the second structure as the initial structure. The above four test proteins sample a relatively wide range of protein size varying from 214 to 819 residues and conformational change with the initial root-mean-square-displacement (RMSD) ranging from 3.77 Å to 13.61 Å.

Table 5.2 Results of structure optimizations for the four benchmark proteins: adenylate kinase, LAO binding protein, maltodextrin-binding protein and elongation factor 2. For adenylate kinase, both the final RMSD and the best RMSD value obtained during the optimization process are shown.

PDB	Initial RMSD (Å)	Scaling factor	Final RMSD (Å)	Final <i>R</i> value
4AKE → 1AKE (214 residues)	7.13	0.06	3.26 3.02(best)	1.36×10^{-4}
2LAO → 1LST (238 residues)	4.70	0.06	1.28	4.28×10^{-6}
1ANF → 1OMP (380 residues)	3.77	0.03	1.63	6.79×10^{-7}
1N0U → 1N0V (819 residues)	13.61	0.015	3.58	2.63×10^{-7}

The results of the structure optimizations are listed in Table 5.2, including the scaling factors used and the RMSD values obtained. The scaling factor c in equation (5.10) ensures the two parts of energy for SAXS fitting and structure restraint are approximately in the same order of magnitude at the beginning of the optimization, so that the optimization is not driven by any specific term only. As the Gō-like model energy of the initial structure scales approximately with the number of residues in the structure, the scaling factor was empirically determined, with the value around 12 divided by the number of residues. It is found that a smaller scaling factor will lead to faster decreasing of RMSD at the beginning of the optimization but will tend to cause over-fitting at large iteration numbers. Nevertheless, the results obtained by different scaling factors have similar final optimized RMSD values in our test (data not shown). It is seen in Table 5.2 that all the four final RMSD values obtained in our structure optimizations show significant improvements compared with their corresponding initial RMSD values. This means the final structures obtained in our structure optimization process align reasonably well with the target structures. A better view of the structure alignment is shown in Figure 5.7, where the cartoon representations of structures before and after optimization are superposed to their corresponding target structures, for all the four benchmark cases. PULCHRA [104] is used to recover atomistic structures from the C α representations for the final optimized structures in order to generate the cartoon representation shown in Figure 5.7.

In Figure 5.8, the evolution of the total optimization energy W , the SAXS-fitting energy W_q and the RMSD value during the structure optimization are plotted for LAO binding protein. The jump in the total optimization energy means that a new optimization cycle with a different λ value is started, where a

larger weight is assigned to the SAXS-fitting energy W_q . From Figure 5.8, it is observed that the RMSD value gradually decreases during the major iteration cycles as the weight of the SAXS-fitting energy is gradually increasing. This resembles the simulated annealing process, where the temperature is gradually decreased. Thus, varying the weight in the optimization gives a better chance finding the correct local minimum, and a similar technique was also implemented in literature [25].

Figure 5.7 Results of structure optimization using our implicit hydration layer to compute scattering intensity for four example proteins. The initial and target structures are shown in white and red respectively. Our optimization result is shown in green. The atomistic structure of our optimization result used to generate the cartoon representation is recovered from the $C\alpha$ coordinates by the software PULCHRA. The four example proteins are (a) adenylate kinase, (b) LAO binding protein, (c) maltodextrin-binding protein and (d) elongation factor 2.

Figure 5.8 The evolution of the total optimization energy (shown in red curve), SAXS-fitting energy (shown in green curve) and RMSD value (shown in blue curve) during our structure optimization process using LAO binding protein as an example. The jump in the total optimization energy means the starting of a new optimization cycle with a larger λ value, where a larger weight is assigned to the SAXS-fitting energy W_q .

It is also noted that, for adenylate kinase, the RMSD value first decreases from 7.13 Å to the minimum value of 3.02 Å and then starts to increase. The final RMSD value of 3.26 Å is slightly larger than the best RMSD value obtained. This indicates over-fitting at the last several iteration cycles. This over-fitting

will occur when the weight of SAXS-fitting energy is very close to 1. So a good criterion to terminate the optimization in time is the key to control the over-fitting issue. It is found that for other three proteins, the RMSD values are constantly decreasing during the whole structure optimization process, like the case shown in Figure 5.8. Thus, the convergence criterion used in our study is designed to deal with different situations, which successfully avoided the over-fitting problem for three out of our four benchmark proteins and works reasonably for all our test cases.

5.7 Conclusions

Despite there are already quite a lot of methods to evaluate the theoretical scattering intensities from protein coordinates, few of them are suitable for performing structure optimization using gradient-based algorithms. This is usually because the hydration layer scattering cannot be directly related to the protein coordinates. In this work, we propose to incorporate an implicit hydration term into the original form factor of protein atom (or coarse-grained bead), in order to represent the contribution of hydration solvent molecules. Our approach is based on the SASA of protein atom (or coarse-grained bead). Using this approach, the analytical derivatives of the scattering intensity with respect to the protein coordinates can be easily obtained. The implicit hydration term is obtained by fitting the theoretical scattering intensities computed using explicit hydration water molecules for all the high resolution structures in our dataset. It is also shown that the structure optimizations with analytical derivatives can be performed with a properly defined pseudo energy function, which contains both a structure-constraint term and a SAXS-fitting term. The structure optimizations

performed using four benchmark proteins can obtain final structures with very small RMSD values compared with the target structures. Our approach of using fitted implicit hydration term to represent the contribution of hydration layer scattering can be of great usage in structure modeling applications based on experimental SAXS profiles. Future directions include the extension of our optimization approach to the case of global searching such as simulated annealing and SAXS-guided MD simulation.

6 Concluding remarks

As a summary, we have proposed improvements to the current methods of evaluating the solution X-ray scattering intensity for both computational efficiency and accuracy. In addition, our scattering intensity computation formula with an implicit hydration term is advantageous in structure modeling.

The FFT-based orientational average method shown in Chapter 3 is suitable for evaluating WAXS or near-WAXS profiles. The structure details of explicit hydration water molecules, usually from explicit MD simulations, are necessary for accurately evaluating the scattering intensity at a high q region. A large number of water molecules will be included in the protein hydration shell for big protein complexes. Thus, when considering the conformational fluctuations from the protein molecule itself by averaging over multiple simulation frames, the computation cost will be prohibited for such large biomolecular systems. Our FFT-based method is suitable for tackle the computation expense issue, as the required computational is nearly a constant for systems with different sizes.

For modeling SAXS profiles, usually a coarse-grained representation is enough as the corresponding resolution of experimental data is relatively low. In this case, the accuracy of the CG approach tends to be more important than the computational efficiency as there are only a small number of CG beads. Our EDM method ensures mathematically the optimal solution for a given CG scheme by using the least squares fitting method. The physical meaning of our EDM-derived form factors considers the correlation between different CG beads and should be more physically reasonable than the widely used IBA approach in literature. Theoretically, the obtained CG form factors are readily

applicable to any CG-based approaches to evaluate the scattering intensity. We show in our work that the obtained CG form factors can be used to fit the experimental SAXS profiles by considering the hydration layer and protein excluded volume corrections.

By including the implicit hydration term fitted in Chapter 5, the computation of scattering intensity becomes analytically differentiable with respect to the protein coordinates. This feature allows a protein structure optimization based on derivatives or a SAXS-biased MD simulation. In our work, we have tested structure optimizations based on theoretical SAXS profiles from known target structures using the conjugate gradient algorithm. It is shown that using a simple structure-constraint potential, the final optimized structures closely resemble the target structures.

Taken together, the above works contribute to the ultimate goal of deriving protein structure models from experimental scattering profiles for large proteins and protein complexes. Future directions include adopting a global search algorithm and developing more sophisticated structure-constraint potential energy function in our structure optimization approach.

References

1. Chu, B. & Hsiao, B. S. (2001) Small-angle X-ray scattering of polymers. *Chem. Rev.* **101**, 1727-1761.
2. Svergun, D. I. & Koch, M. H. J. (2003) Small-angle scattering studies of biological macromolecules in solution. *Rep. Prog. Phys.* **66**, 1735-1782.
3. Doniach, S. (2001) Changes in biomolecular conformation seen by small angle X-ray scattering. *Chem. Rev.* **101**, 1763-1778.
4. Koch, M. H. J., Vachette, P. & Svergun, D. I. (2003) Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Q. Rev. Biophys.* **36**, 147-227.
5. Lipfert, J. & Doniach, S. (2007) Small-angle X-ray scattering from RNA, proteins, and protein complexes. *Annu. Rev. Bioph. Biom.* **36**, 307-327.
6. Petoukhov, M. V. & Svergun, D. I. (2007) Analysis of X-ray and neutron scattering from biomacromolecular solutions. *Curr. Opin. Struc. Biol.* **17**, 562-571.
7. Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. A. (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q. Rev. Biophys.* **40**, 191-285.
8. Toft, K. N., Vestergaard, B., Nielsen, S. S., Snakenborg, D., Jeppesen, M. G., Jacobsen, J. K., Arleth, L. & Kutter, J. P. (2008) High-throughput Small Angle X-ray Scattering from proteins in solution using a microfluidic front-end. *Anal. Chem.* **80**, 3648-3654.
9. 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) Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat. Methods* **6**, 606-U683.

10. Mertens, H. D. T. & Svergun, D. I. (2010) Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *J. Struct. Biol.* **172**, 128-141.
11. Graewert, M. A. & Svergun, D. I. (2013) Impact and progress in small and wide angle X-ray scattering (SAXS and WAXS). *Curr. Opin. Struc. Biol.* **23**, 748-754.
12. Blanchet, C. E. & Svergun, D. I. (2013) Small-Angle X-Ray Scattering on Biological Macromolecules and Nanocomposites in Solution. *Annu. Rev. Phys. Chem.* **64**, 37-54.
13. Svergun, D. I. & Koch, M. H. J. (2002) Advances in structure analysis using small-angle scattering in solution. *Curr. Opin. Struc. Biol.* **12**, 654-660.
14. Putnam, D. K., Lowe, E. W., Jr. & Meiler, J. (2013) Reconstruction of SAXS Profiles from Protein Structures. *Comput Struct Biotechnol J* **8**, e201308006.
15. Valentini, E., Kikhney, A. G., Previtali, G., Jeffries, C. M. & Svergun, D. I. (2015) SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Res.* **43**, D357-D363.
16. Debye, P. (1915) Scattering of x-rays [Zerstreuung von Ro'ntgenstrahlen]. *Ann Phys-Berlin* **351**, 809-823.
17. Svergun, D. I., Richard, S., Koch, M. H. J., Sayers, Z., Kuprin, S. & Zaccai, G. (1998) Protein hydration in solution: Experimental observation by x-ray and neutron scattering. *Proc. Natl. Acad. Sci. USA* **95**, 2267-2272.
18. Ebbinghaus, S., Kim, S. J., Heyden, M., Yu, X., Heugen, U., Gruebele, M., Leitner, D. M. & Havenith, M. (2007) An extended dynamical hydration shell around proteins. *Proc. Natl. Acad. Sci. USA* **104**, 20749-20752.
19. Zhang, L. Y., Wang, L. J., Kao, Y. T., Qiu, W. H., Yang, Y., Okobiah, O. & Zhong, D. P. (2007) Mapping hydration dynamics around a protein surface. *Proc. Natl. Acad. Sci. USA* **104**, 18461-18466.

20. Svergun, D., Barberato, C. & Koch, M. H. J. (1995) CRY SOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* **28**, 768-773.
21. Schneidman-Duhovny, D., Hammel, M. & Sali, A. (2010) FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res.* **38**, W540-W544.
22. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. (2013) Accurate SAXS Profile Computation and its Assessment by Contrast Variation Experiments. *Biophys. J.* **105**, 962-974.
23. Tjioe, E. & Heller, W. T. (2007) ORNL_SAS: software for calculation of small-angle scattering intensities of proteins and protein complexes. *J. Appl. Crystallogr.* **40**, 782-785.
24. Bardhan, J., Park, S. & Makowski, L. (2009) SoftWAXS: a computational tool for modeling wide-angle X-ray solution scattering from biomolecules. *J. Appl. Crystallogr.* **42**, 932-943.
25. Zheng, W. J. & Tekpinar, M. (2011) Accurate Flexible Fitting of High-Resolution Protein Structures to Small-Angle X-Ray Scattering Data Using a Coarse-Grained Model with Implicit Hydration Shell. *Biophys. J.* **101**, 2981-2991.
26. Liu, H. G., Morris, R. J., Hexemer, A., Grandison, S. & Zwart, P. H. (2012) Computation of small-angle scattering profiles with three-dimensional Zernike polynomials. *Acta Crystallogr. A* **68**, 278-285.
27. Virtanen, J. J., Makowski, L., Sosnick, T. R. & Freed, K. F. (2011) Modeling the Hydration Layer around Proteins: Applications to Small- and Wide-Angle X-Ray Scattering. *Biophys. J.* **101**, 2061-2069.
28. Poitevin, F., Orland, H., Doniach, S., Koehl, P. & Delarue, M. (2011) AquaSAXS: a web server for computation and fitting of SAXS profiles with non-uniformly hydrated atomic models. *Nucleic Acids Res.* **39**, W184-W189.
29. Yang, S. C., Park, S., Makowski, L. & Roux, B. (2009) A Rapid Coarse Residue-Based Computational Method for X-Ray Solution Scattering

Characterization of Protein Folds and Multiple Conformational States of Large Protein Complexes. *Biophys. J.* **96**, 4449-4463.

30. Ravikumar, K. M., Huang, W. & Yang, S. C. (2013) Fast-SAXS-Pro: A Unified Approach to Computing SAXS Profiles of DNA, RNA, Protein, and their Complexes. *Biophys. J.* **104**, 183a-183a.
31. Grishaev, A., Guo, L. A., Irving, T. & Bax, A. (2010) Improved Fitting of Solution X-ray Scattering Data to Macromolecular Structures and Structural Ensembles by Explicit Water Modeling. *J. Am. Chem. Soc.* **132**, 15484-15486.
32. Park, S., Bardhan, J. P., Roux, B. & Makowski, L. (2009) Simulated x-ray scattering of protein solutions using explicit-solvent models. *J. Chem. Phys.* **130**, 134114.
33. Chen, P. C. & Hub, J. S. (2014) Validating Solution Ensembles from Molecular Dynamics Simulation by Wide-Angle X-ray Scattering Data. *Biophys. J.* **107**, 435-447.
34. Oroguchi, T., Hashimoto, H., Shimizu, T., Sato, M. & Ikeguchi, M. (2009) Intrinsic Dynamics of Restriction Endonuclease EcoO109I Studied by Molecular Dynamics Simulations and X-Ray Scattering Data Analysis. *Biophys. J.* **96**, 2808-2822.
35. Kofinger, J. & Hummer, G. (2013) Atomic-resolution structural information from scattering experiments on macromolecules in solution. *Phys. Rev. E* **87**, 052712.
36. Fraser, R. D. B., Macrae, T. P. & Suzuki, E. (1978) Improved Method for Calculating Contribution Of Solvent To X-Ray-Diffraction Pattern Of Biological Molecules. *J. Appl. Crystallogr.* **11**, 693-694.
37. Stuhrman.Hb (1970) Interpretation Of Small-Angle Scattering Functions Of Dilute Solutions And Gases - a Representation Of Structures Related To a One-Particle-Scattering Function. *Acta Crystall a-Crys A* **26**, 297-306.
38. Koehl, P. & Delarue, M. (2010) AQUASOL: An efficient solver for the dipolar Poisson-Boltzmann-Langevin equation. *J. Chem. Phys.* **132**, 064101.

39. Luchko, T., Gusarov, S., Roe, D. R., Simmerling, C., Case, D. A., Tuszynski, J. & Kovalenko, A. (2010) Three-Dimensional Molecular Theory of Solvation Coupled with Molecular Dynamics in Amber. *J. Chem. Theory Comput.* **6**, 607-624.
40. Tong, D., Yang, J. & Lu, L. (2015) Modeling solution X-ray scattering of biomacromolecules using an explicit solvent model and the fast Fourier transform. *J. Appl. Crystallogr.* **48**, 1834-1842.
41. Niebling, S., Björling, A. & Westenhoff, S. (2014) MARTINI bead form factors for the analysis of time-resolved X-ray scattering of proteins. *J. Appl. Crystallogr.* **47**, 1190-1198.
42. Stovgaard, K., Andreetta, C., Ferkinghoff-Borg, J. & Hamelryck, T. (2010) Calculation of accurate small angle X-ray scattering curves from coarse-grained protein models. *Bmc Bioinformatics* **11**, 429.
43. Svergun, D. I. (1999) Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* **76**, 2879-2886.
44. Franke, D. & Svergun, D. I. (2009) DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* **42**, 342-346.
45. Gorba, C., Miyashita, O. & Tama, F. (2008) Normal-mode flexible fitting of high-resolution structure of biological molecules toward one-dimensional low-resolution data. *Biophys. J.* **94**, 1589-1599.
46. Panjkovich, A. & Svergun, D. I. (2016) Deciphering conformational transitions of proteins by small angle X-ray scattering and normal mode analysis. *Phys Chem Chem Phys* **18**, 5707-5719.
47. Svergun, D. I., Petoukhov, M. V. & Koch, M. H. J. (2001) Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.* **80**, 2946-2953.
48. Björling, A., Niebling, S., Marcellini, M., van der Spoel, D. & Westenhoff, S. (2015) Deciphering Solution Scattering Data with Experimentally Guided Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **11**, 780-787.

49. Kimanius, D., Pettersson, I., Schluckebier, G., Lindahl, E. & Andersson, M. (2015) SAXS-Guided Metadynamics. *J. Chem. Theory Comput.* **11**, 3491-3498.
50. Chen, P. C. & Hub, J. S. (2015) Interpretation of Solution X-Ray Scattering by Explicit-Solvent Molecular Dynamics. *Biophys. J.* **108**, 2573-2584.
51. Schmidt-Rohr, K. (2007) Simulation of small-angle scattering curves by numerical Fourier transformation. *J. Appl. Crystallogr.* **40**, 16-25.
52. Brown, P. J. F., A. G.; Maslen, E. N.; O'Keefe, M. A.; Willis, B. T. M. (2006). Vol. C. *International Tables for Crystallography*, edited by P. E., pp. 554-595. Berlin, Heidelberg, Germany: Springer.
53. Wang, G. L. & Dunbrack, R. L. (2003) PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589-1591.
54. Sali, A. & Blundell, T. L. (1993) Comparative Protein Modeling by Satisfaction Of Spatial Restraints. *J. Mol. Biol.* **234**, 779-815.
55. Pronk, S., Pall, 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) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845-854.
56. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586-3616.
57. Jayaram, B., Bhushan, K., Shenoy, S. R., Narang, P., Bose, S., Agrawal, P., Sahu, D. & Pandey, V. (2006) Bhageerath: an energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins. *Nucleic Acids Res.* **34**, 6195-6204.

58. Castiglioni, P. (2014). *Wiley StatsRef: Statistics Reference Online*: John Wiley & Sons, Ltd.
59. Diamond, R. (1974) Real-space refinement of the structure of hen egg-white lysozyme. *J Mol Biol* **82**, 371-391.
60. Maurus, R., Overall, C. M., Bogumil, R., Luo, Y., Mauk, A. G., Smith, M. & Brayer, G. D. (1997) A myoglobin variant with a polar substitution in a conserved hydrophobic cluster in the heme binding pocket. *Bba-Protein Struct. M.* **1341**, 1-13.
61. Bell, J. A. (1999) X-ray crystal structures of a severely desiccated protein. *Protein Sci.* **8**, 2033-2040.
62. Hekkelman, M. L., Beek, T. A. H. T., Pettifer, S. R., Thorne, D., Attwood, T. K. & Vriend, G. (2010) WIWS: a protein structure bioinformatics Web service collection. *Nucleic Acids Res.* **38**, W719-W723.
63. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983) Comparison Of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **79**, 926-935.
64. Golub, G. & Kahan, W. (1965) Calculating the Singular Values and Pseudo-Inverse of a Matrix. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis* **2**, 205-224.
65. Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. (1987) The Missing Term In Effective Pair Potentials. *J. Phys. Chem.* **91**, 6269-6271.
66. Sorenson, J. M., Hura, G., Glaeser, R. M. & Head-Gordon, T. (2000) What can x-ray scattering tell us about the radial distribution functions of water? *J. Chem. Phys.* **113**, 9149-9161.
67. Nelder, J. A. & Mead, R. (1965) A Simplex-Method for Function Minimization. *Comput. J.* **7**, 308-313.
68. Bayer, P., Arndt, A., Metzger, S., Mahajan, R., Melchior, F., Jaenicke, R. & Becker, J. (1998) Structure determination of the small ubiquitin-related modifier SUMO-1. *J. Mol. Biol.* **280**, 275-286.

69. De Guzman, R. N., Martinez-Yamout, M. A., Dyson, H. J. & Wright, P. E. (2004) Interaction of the TAZ1 domain of the CREB-binding protein with the activation domain of CITED2: regulation by competition between intrinsically unstructured ligands for non-identical binding sites. *J. Biol. Chem.* **279**, 3042-3049.
70. Song, J., Lee, M. S., Carlberg, I., Vener, A. V. & Markley, J. L. (2006) Micelle-induced folding of spinach thylakoid soluble phosphoprotein of 9 kDa and its functional implications. *Biochemistry* **45**, 15633-15643.
71. Lee, C. W., Martinez-Yamout, M. A., Dyson, H. J. & Wright, P. E. (2010) Structure of the p53 transactivation domain in complex with the nuclear receptor coactivator binding domain of CREB binding protein. *Biochemistry* **49**, 9964-9971.
72. Singarapu, K. K., Tonelli, M., Chow, D. C., Frederick, R. O., Westler, W. M. & Markley, J. L. (2011) Structural characterization of Hsp12, the heat shock protein from *Saccharomyces cerevisiae*, in aqueous solution where it is intrinsically disordered and in detergent micelles where it is locally alpha-helical. *The Journal of biological chemistry* **286**, 43447-43453.
73. Rowell, J. P., Simpson, K. L., Stott, K., Watson, M. & Thomas, J. O. (2012) HMGB1-facilitated p53 DNA binding occurs via HMG-Box/p53 transactivation domain interaction, regulated by the acidic tail. *Structure* **20**, 2014-2024.
74. Fei, X., Ye, X., LaRonde, N. A. & Lorimer, G. H. (2014) Formation and structures of GroEL:GroES2 chaperonin footballs, the protein-folding functional form. *Proc. Natl. Acad. Sci. USA* **111**, 12775-12780.
75. Wright, D. W. & Perkins, S. J. (2015) SCT: a suite of programs for comparing atomistic models with small-angle scattering data. *J. Appl. Crystallogr.* **48**, 953-961.
76. Li, Y. Q. & Zhang, Y. (2009) REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins* **76**, 665-674.
77. Wassenaar, T. A., Pluhackova, K., Bockmann, R. A., Marrink, S. J. & Tieleman, D. P. (2014) Going Backward: A Flexible Geometric

Approach to Reverse Transformation from Coarse Grained to Atomistic Models. *J. Chem. Theory Comput.* **10**, 676-690.

78. Reith, D., Putz, M. & Muller-Plathe, F. (2003) Deriving effective mesoscale potentials from atomistic simulations. *J Comput Chem* **24**, 1624-1636.
79. Lyubartsev, A. P. & Laaksonen, A. (1995) Calculation Of Effective Interaction Potentials From Radial-Distribution Functions - a Reverse Monte-Carlo Approach. *Phys. Rev. E* **52**, 3730-3737.
80. Izvekov, S. & Voth, G. A. (2005) Multiscale coarse graining of liquid-state systems. *J. Chem. Phys.* **123**, 134105.
81. Noid, W. G., Chu, J. W., Ayton, G. S., Krishna, V., Izvekov, S., Voth, G. A., Das, A. & Andersen, H. C. (2008) The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **128**, 244114.
82. Miyazawa, S. & Jernigan, R. L. (1985) Estimation Of Effective Interresidue Contact Energies From Protein Crystal-Structures - Quasi-Chemical Approximation. *Macromolecules* **18**, 534-552.
83. Sippl, M. J. (1995) Knowledge-Based Potentials for Proteins. *Curr. Opin. Struc. Biol.* **5**, 229-235.
84. Jernigan, R. L. & Bahar, I. (1996) Structure-derived potentials and protein simulations. *Curr. Opin. Struc. Biol.* **6**, 195-209.
85. Miyazawa, S. & Jernigan, R. L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623-644.
86. Cooper, C. D. O., Newman, J. A., Aitkenhead, H., Allerston, C. K. & Gileadi, O. (2015) Structures of the Ets Protein DNA-binding Domains of Transcription Factors Etv1, Etv4, Etv5, and Fev DETERMINANTS OF DNA BINDING AND REDOX REGULATION BY DISULFIDE BOND FORMATION. *J. Biol. Chem.* **290**, 13692-13709.
87. 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) The energy landscape of adenylate kinase during catalysis. *Nat. Struct. Mol. Biol.* **22**, 124-131.
88. Dobbek, H., Gremer, L., Kiefersauer, R., Huber, R. & Meyer, O. (2002) Catalysis at a dinuclear [CuSMo(=O)OH] cluster in a CO dehydrogenase resolved at 1.1-angstrom resolution. *Proc. Natl. Acad. Sci. USA* **99**, 15971-15976.
 89. Vijaykumar, S., Bugg, C. E. & Cook, W. J. (1987) Structure Of Ubiquitin Refined at 1.8 a Resolution. *J. Mol. Biol.* **194**, 531-544.
 90. Bushnell, G. W., Louie, G. V. & Brayer, G. D. (1990) High-resolution three-dimensional structure of horse heart cytochrome c. *J Mol Biol* **214**, 585-595.
 91. Wang, D. C., Bode, W. & Huber, R. (1985) Bovine Chymotrypsinogen-a X-Ray Crystal-Structure Analysis And Refinement Of a New Crystal Form at 1.8 a Resolution. *J. Mol. Biol.* **185**, 595-624.
 92. Tong, D., Yang, S. & Lu, L. (2016) Accurate optimization of amino acid form factors for computing small-angle X-ray scattering intensity of atomistic protein structures. *J. Appl. Crystallogr.* **49**.
 93. Cavallo, L., Kleinjung, J. & Fraternali, F. (2003) POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Res.* **31**, 3364-3366.
 94. Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000) Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **298**, 937-953.
 95. Noel, J. K., Whitford, P. C. & Onuchic, J. N. (2012) The Shadow Map: A General Contact Definition for Capturing the Dynamics of Biomolecular Folding and Function. *J. Phys. Chem. B* **116**, 8692-8702.
 96. Hasel, W., Hendrickson, T. F. & Still, W. C. (1988) A rapid approximation to the solvent accessible surface areas of atoms. *Tetrahedron Computer Methodology* **1**, 103-116.

97. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical recipes in C (2nd ed.): the art of scientific computing*. Cambridge University Press.
98. Cromer, D. T. & Waber, J. T. (1965) Scattering Factors Computed From Relativistic Dirac-Slater Wave Functions. *Acta Crystallogr.* **18**, 104-109.
99. Muller, C. W. & Schulz, G. E. (1992) Structure Of the Complex Between Adenylate Kinase From Escherichia-Coli And the Inhibitor Ap5a Refined at 1.9 a Resolution - a Model for a Catalytic Transition-State. *J. Mol. Biol.* **224**, 159-177.
100. Muller, C. W., Schlauderer, G. J., Reinstein, J. & Schulz, G. E. (1996) Adenylate kinase motions during catalysis: An energetic counterweight balancing substrate binding. *Structure* **4**, 147-156.
101. Oh, B. H., Pandit, J., Kang, C. H., Nikaido, K., Gokcen, S., Ames, G. F. & Kim, S. H. (1993) Three-dimensional structures of the periplasmic lysine/arginine/ornithine-binding protein with and without a ligand. *J. Biol. Chem.* **268**, 11348-11355.
102. Sharff, A. J., Rodseth, L. E., Spurlino, J. C. & Quijcho, F. A. (1992) Crystallographic Evidence Of a Large Ligand-Induced Hinge-Twist Motion Between the 2 Domains Of the Maltodextrin Binding-Protein Involved In Active-Transport And Chemotaxis. *Biochemistry* **31**, 10657-10663.
103. Jorgensen, R., Ortiz, P. A., Carr-Schmid, A., Nissen, P., Kinzy, T. G. & Andersen, G. R. (2003) Two crystal structures demonstrate large conformational changes in the eukaryotic ribosomal translocase. *Nat Struct Biol* **10**, 379-385.
104. Rotkiewicz, P. & Skolnick, J. (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. *J Comput Chem* **29**, 1460-1465.

Publications

1. Xia, F.[†], **Tong, D. D.**[†] & Lu, L. Y. (2013) Robust Heterogeneous Anisotropic Elastic Network Model Precisely Reproduces the Experimental B-factors of Biomolecules. *J. Chem. Theory Comput.* **9**, 3704-3714.

[†] *These two authors contribute equally to the work.*

2. Xia, F., **Tong, D. D.**, Yang, L. F., Wang, D. Y., Hoi, S. C. H., Koehl, P. & Lu, L. Y. (2014) Identifying Essential Pairwise Interactions in Elastic Network Model Using the Alpha Shape Theory. *J Comput. Chem.* **35**, 1111-1121.

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

4. **Tong, D. D.**, Yang, S. C. & Lu, L. Y. (2016) Accurate optimization of amino acid form factors for computing small-angle X-ray scattering intensity of atomistic protein structures. *J. Appl. Crystallogr.* **49**.

5. **Tong, D. D.**, Yang, S. C. & Lu, L. Y. (2016) Modeling protein hydration shell using coarse-grained scattering form factors. *J. Appl. Crystallogr.* (Submitted)