A Rapid Coarse Residue-Based Computational Method for X-Ray Solution Scattering Characterization of Protein Folds and Multiple Conformational States of Large Protein Complexes

Sichun Yang,[†] Sanghyun Park,[‡] Lee Makowski,[§] and Benoît Roux[†§*]

[†]Department of Biochemistry and Molecular Biology, Gordon Center for Integrative Science, The University of Chicago, Chicago, Illinois; and [‡]Mathematics and Computer Science Division and [§]Biosciences Division, Argonne National Laboratory, Argonne, Illinois

ABSTRACT We present a coarse residue-based computational method to rapidly compute the solution scattering profile from a protein with dynamical fluctuations. The method is built upon a coarse-grained (CG) representation of the protein. This CG representation takes advantage of the intrinsic low-resolution and CG nature of solution scattering data. It allows rapid scattering determination from a large number of conformations that can be extracted from CG simulations to obtain scattering characterization of protein conformations. The method includes several important elements, effective residue structure factors derived from the Protein Data Bank, explicit treatment of water molecules in the hydration layer at the surface of the protein, and an ensemble average of scattering from a variety of appropriate conformations to account for macromolecular flexibility. This simplified method is calibrated and illustrated to accurately reproduce the experimental scattering curve of Hen egg white lysozyme. We then illustrated the applications of this CG method by computing the solution scattering patterns of several representative protein folds and multiple conformational states. The results suggest that solution scattering data, when combined with the reliable computational method that we developed, show great potential for a better structural description of multidomain complexes in different functional states, and for recognizing structural folds when sequence similarity to a protein of known structure is low.

INTRODUCTION

Small-angle x-ray solution scattering (SAXS) is an increasingly powerful technique to characterize structurally large macromolecular complexes (1–6). It takes less effort for sample preparation relative to crystallography, and avoids the challenge of growing crystals of good diffraction quality. It provides native structural data at physiological conditions such as in NMR, but without the inherent size limitations. It also allows rapid data collection with the current high flux synchrotron sources. The tradeoff is that only low-resolution information (in the range of 10–50 Å; see (1) for more discussion about SAXS resolution) about overall shape can be obtained because of the spherical averaging of protein scattering from multiple random orientations adopted in solution.

Low-resolution SAXS data can be combined with computational methods to reconstruct low-resolution structural models of large multidomain complexes. The scattering data can serve as independent constraints for computational modeling to ultimately characterize the structures/shapes of large protein complexes, especially when the structure of each individual domain in the complex is known at high resolution (1). This combination of solution scattering with computation and atomic resolution structures from crystallography provides an alternative approach to achieving structural characterization of multidomain proteins (1). Recent studies suggest that such a combination can be used to obtain the structural multiplicity of multidomain complexes in solution (7,8). However, this application is often limited by the efficiency of the scattering calculations for a large number of conformations generated from extensive sampling in the configurational space. Before this combination can be achieved, we need to develop a rapid scattering determination method to efficiently and accurately compute the scattering profiles so that it can be productively applied to an ensemble of protein states, e.g., tens of thousands of protein configurations.

Recently, several computational approaches have been introduced to address this question, at both the all-atom and coarse-grain levels of detail. In most current all-atom methods, the details of both the protein itself and the surrounding water molecules are approximately taken into account (9–12). One of the most widely used all-atom methods is provided by the CRYSOL program (9), which has been shown to be quite successful in many cases (13). The current treatment in CRYSOL makes assumptions about the structure of primary hydration shell around the protein using an implicit solvent model and about the electron density in the shell. Additionally, the treatment of the density contrast of bound waters in the hydration shell relative to the remaining bulk water remains uncertain (14–17). For large macromolecules, the conformational flexibility must be also taken into account. This flexibility is an important aspect of multidomain complexes and enzymes (18,19) and can be reflected in dynamic transitions among those states accessible to a protein in solution (20). Complete and accurate computation of SAXS patterns would require that it be included in any model.

Submitted September 29, 2008, and accepted for publication March 4, 2009.

*Correspondence: roux@uchicago.edu

Editor: Angel E. Garcia.

0006-3495/09/06/4449/15 $2.00

doi: 10.1016/j.bpj.2009.03.036

Because atomic details are included, these all-atom calculations are often computationally expensive for large complexes, especially when a large number of configurations is involved. Alternative methods for computing scattering patterns involve coarse-graining molecular representations. These are essentially based on the nature of SAXS as a low-resolution technique, making it well suited for use with coarse-graining. Ideally, it should significantly reduce computer time without compromising accuracy. Along this direction, multiple levels of coarse-grained (CG) models have been introduced, including a simple Cα model (21), a side-chain CG model (primarily for diffraction) (22), and a dummy-residue (DR) model (23). In the Cα model, all amino acids are assigned with the same number of electrons at the Cα positions (21,24). This simple Cα model approach has reportedly improved the prediction ability of protein folding and structure prediction (for example, see (25,26)). In the side-chain coarse-graining model, a procedure of simplifying each residue into backbone and side-chain groups was aimed to obtain low-resolution interpretation for diffraction data (22,27,28). In the DR model, each residue is represented by its Cα atom and explicit water molecules are used in the hydration shell (23). In addition, the use of a correction function for the structure factor of each residue was enforced to reproduce the CRYSOL-calculated scattering curves. Furthermore, in the DR treatment of the hydration shell, the density of explicit water molecules used to represent the hydration shell density is much lower than that of the bulk solvent.

Inspired by the successes of CRYSOL and the DR model, we aim here at developing a CG model that can be used to efficiently and accurately compute scattering curves from a given protein conformation. In this CG model, three major elements are addressed as follows.

First, a knowledge-based structure factor for each residue is developed based on their atomic models as in the Protein Data Bank (PDB). This differs from the Cα model by including a CG structure factor for each residue. It also differs from the DR model by avoiding the introduction of a correction function since the effective structure factors are essentially derived from experimentally observed conformers. Clearly, this coarse-graining is based on the low-resolution nature of SAXS data.

Second, an explicit solvent layer of dummy water molecules is placed around the protein, similar to the DR model, but with a proper electron density set to account for the excess electron density of the hydration layer relative to bulk solvent. An effective structure factor of dummy waters is derived and a proper weight is assigned based on the experimental data of lysozyme.

Third, we use molecular dynamics (MD) simulations to account for the conformational flexibility that occurs in proteins in solution. The CG model is capable of reproducing the SAXS scattering data of lysozyme accurately. More importantly, it provides a rapid determination method for computing scattering profiles with an ensemble of states incorporated. We further apply the CG model to characterize a variety of protein folds and multiple conformational states in terms of their distinct scattering profiles. This rapid computational method, when combined with solution scattering data and atomic resolution structures of individual components, is well positioned to provide a powerful tool to shape-reconstruct large multidomain complexes and to determine the population fraction of different conformational states of protein under different physiological conditions.

Theoretical background and developments

The SAXS from protein in solution essentially measures the difference in the electron density between protein molecules and bulk solvent. Proteins have an average electron density of $\rho_m \sim 0.43\ e/\text{Å}^3$ (29), whereas pure water has an electron density of $\rho_s \sim 0.334\ e/\text{Å}^3$ at 20°C (30). The difference makes SAXS particularly attractive for resolving the electron density contrast and potentially determining protein structures. In practice, the scattering curve $I(q)$ is determined from the total scattering of protein samples by subtracting the capillary scattering and background buffer scattering after correction for the volume of buffer displaced by the protein (1,31). Theoretically, the intensity from dilute samples is proportional to the spherically averaged scattering of a single molecule minus the excluded volume contributions but with the hydration shell excess density (3),

$$I(q) = \left\langle |A_m(\mathbf{q}) - \rho_s A_s(\mathbf{q}) + \Delta\rho_b A_b(\mathbf{q})|^2 \right\rangle_\Omega, \quad (1)$$

where the amplitude of the wavevector transfer $q = |\mathbf{q}| = 2\pi/d = 4\pi \sin\theta/\lambda$ (d is the Bragg spacing, 2θ is the scattering angle, and λ is the x-ray wavelength). $A_m(\mathbf{q})$ is the scattering amplitude from the protein molecule in vacuum, $A_s(\mathbf{q})$ is from the solvent with an excluded volume displaced by the protein, and $A_b(\mathbf{q})$ is from the shell of bound waters reflected in the density excess ($\Delta\rho_b$) relative to the bulk phase (29). The quantity $\langle \bullet\bullet\bullet \rangle_\Omega$ stands for an average over all orientations in reciprocal space to account for the nature of protein adopting random orientations in solution.

Equation 1 provides the theoretical basis for solution scattering. For a given protein conformation with N spherical atoms (e.g., Fig. 1), the scattering is contributed from 1), the protein itself in vacuum; 2), the excluded volume of solvent displaced by the protein; 3), the relative electron density in the hydration shell; and 4), in addition, protein conformational flexibility by the ensemble average over all conformations that are accessible to protein motions in solution. The ensemble average differs from the orientational average in Eq. 1. We briefly describe these four aspects as follows. First, the scattering $I(q)$ from the protein itself is calculated by the Debye formula

$$I(q) = \left\langle |A_m(\mathbf{q})|^2 \right\rangle = \sum_{i,j=1}^{N} f_i(q) f_j(q) \frac{\sin(qr_{ij})}{qr_{ij}}, \quad (2)$$

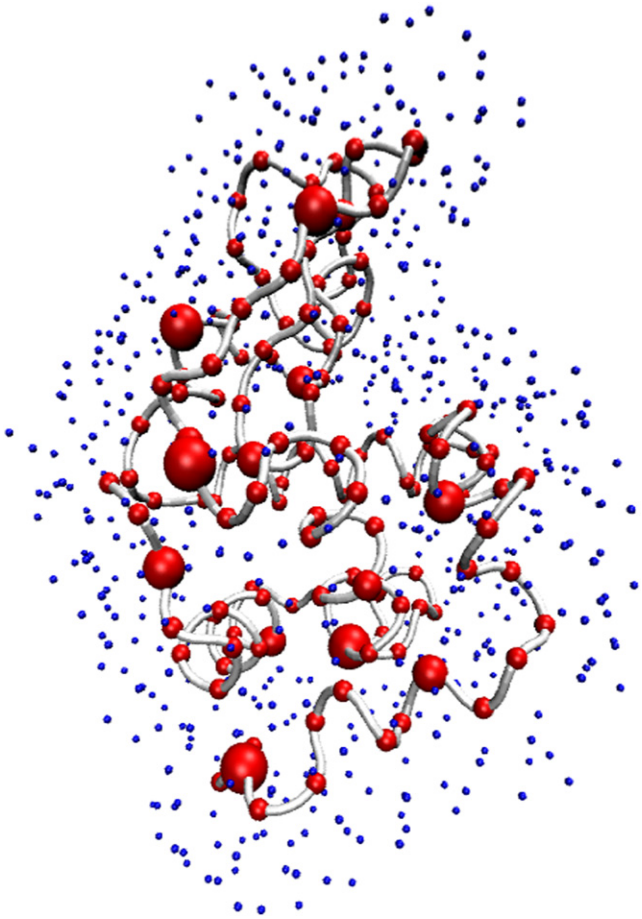

FIGURE 1 Schematic representation of lysozyme surrounded by a layer of "dummy" waters to model the local difference in relative solvent density at the surface of the protein. The scattering from a given protein conformation is conveniently and accurately represented by its N residues (*red*) via the Cα position, with M explicit water molecules (*blue*) inserted 3.5–6.5 Å away from Cα atoms to represent the hydration shell. Conformational flexibility also contributes to the scattering because of the intrinsic motions that are accessible to protein dynamics in solution.

where f_i values are atomic form factors ($i = 1, \cdots, N$) and r_{ij} values are the interparticle distances. At the limit of $q \rightarrow 0$, f_i values are the electron numbers of each atom. Fig. 2 shows the scattering factor curves f_i for C, N, O, H, and S using the Cromer-Mann scattering-factor coefficients (32).

Second, the effect of the excluded solvent can be incorporated into a correction for atomic scattering factors by assigning a Gaussian sphere for all the atoms (30),

$$f_i'(q) = f_i(q) - v_i\rho_s\exp\left(-\pi v_i^{2/3}q^2\right), \qquad (3)$$

where v_i values are the observed volumes of each atom from experiments (30). This treatment has been implemented in the CRYSOL package (9). Therefore, the scattering from the protein taking into account the excluded volume is given by

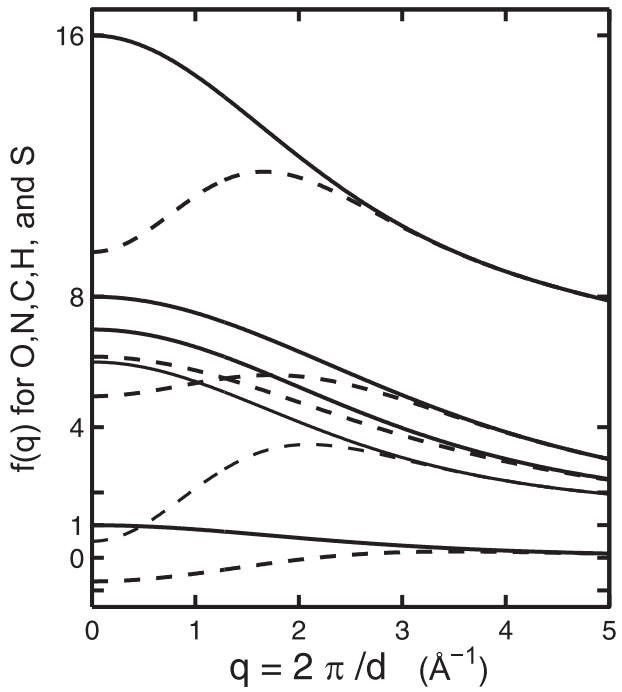

FIGURE 2 Atomic scattering form factors for atoms C, N, O, H, and S, before (*solid line*) and after (*dashed line*) the excluded volume correction. The effect of the excluded volume can be simulated by supposing that the volume displayed by the protein is filled with an electron gas with a density equal to the average electron density of pure water. This has been formulated by assigning a Gaussian sphere for each atom (30), according to Eq. 3.

$$I(q) = \left\langle |A_m(\mathbf{q}) - \rho_s A_s(\mathbf{q})|^2 \right\rangle = \sum_{i,j=1}^{N} f_i'(q) f_j'(q) \frac{\sin(qr_{ij})}{qr_{ij}}, \qquad (4)$$

where $f_i'(q)$ values are the corrected scattering factors in Eq. 3 and plotted in Fig. 2.

Third, the scattering is also contributed from the relative electron density of the primary solvation layer surrounding the protein to the extent that it differs from bulk solvent. It has been documented that the density of bound waters in the hydration shell is slightly higher relative to bulk water (15–17,29). The difference in density gives rise to the third term in the scattering $I(q)$ (Eq. 1). In the CRYSOL calculations, the hydration shell is accounted for by using a water density 10% greater than bulk solvent by default. The level of contrast in the hydration shell can be adjusted to improve the fit to data. The validity of such a representation of the structure of the implicit solvent shell is uncertain (17,33), and therefore an explicit solvent representation will be implemented here.

Finally, the scattering is also affected by the protein conformational flexibility that reflects the nature of protein motions in solution. Modeling such flexibility is fundamentally important to describe the motions that are accessible to molecules outside of a crystal lattice (1). From a theoretical

standpoint, we shall address this issue by adopting MD simulations.

Based on these four theoretical aspects of solution scattering, we have developed a new set of programs for computing solution scattering from atomic coordinates. In these new programs, the scattering intensity is computed by coarse-graining the protein representation with effective residue-based scattering factors, and coarse-graining the protein motions for conformational flexibility using MD simulations. The coarse-graining methods allow us to achieve a significant reduction of computer time, and the large time-scale protein motions are required to adequately reflect the nature of protein dynamics in solution. These programs perform just as well on the test of lysozyme as CRYSOL, and make better assumptions about the solvent density in the primary hydration shell by including the solvent explicitly in the calculations.

Derivation of coarse residue structure factors

Solution scattering has been traditionally used to characterize the protein shape at low resolution (1). Calculations of such low resolution scattering profiles can be accommodated by simplifying the protein representation. We represent the protein as a chain of effective residues specified by the Cα position. Coarse-graining the protein representation is computationally advantageous, though one has to be careful in replacing the atomic scattering factors by effective residue-based structure factors to account for the internal detail of each amino acid (22,28). For each amino acid, an effective structure factor is derived from atomic coordinates of its n spherical atoms using the Debye formula (34)

$$F^{\mathrm{CG}}(q) = \left\langle \sum_{i,j=1}^{\mathrm{n}} f_i'(q) f_j'(q) \frac{\sin(qr_{ij})}{qr_{ij}} \right\rangle_{\mathrm{PDB}}^{\frac{1}{2}}, \qquad (5)$$

where $f_i'(q)$ values are the scattering factors, corrected for excluded volume (Eq. 3 and Fig. 2). The procedure of simplifying a group of spherical atoms into a ''glob'' for each residue is illustrated in the Appendix. This idea of residue-based coarse-graining bears some similarity with the side-chain ''globbicity,'' introduced by Harker (27) and extended by Guo et al. (22,28). The brackets $\langle \cdots \rangle_{\mathrm{PDB}}$ indicates the scattering factor was averaged over backbone conformers and side-chain rotamers of each residue in a set of high-resolution crystal structures of 434 protein chains selected from the PDB (as of July 2008) using the PISCES program (35).

A layer of explicit "dummy" waters

To represent the bound waters in the hydration shell, a layer of explicit water molecules are placed around the surface of protein. For the water molecule, similar to the procedure for amino acids, an effective scattering factor is derived and plotted in Fig. 3 according to

$$F_{\mathrm{w}}^{\mathrm{CG}}(q) = \left[\sum_{i,j=1}^{3} f_i(q) f_j(q) \frac{\sin(qr_{ij})}{qr_{ij}} \right]^{\frac{1}{2}}, \qquad (6)$$

where r_{ij} values are the internal distances taken from a TIP3P model water (36). In practice, such a layer of ''dummy'' waters were placed at their Oxygen positions in a density of bulk solvent (ρ_s) with positions 3.5–6.5 Å away from the protein Cα atoms (Fig. 1), generated using a large equilibrated TIP3P waterbox.

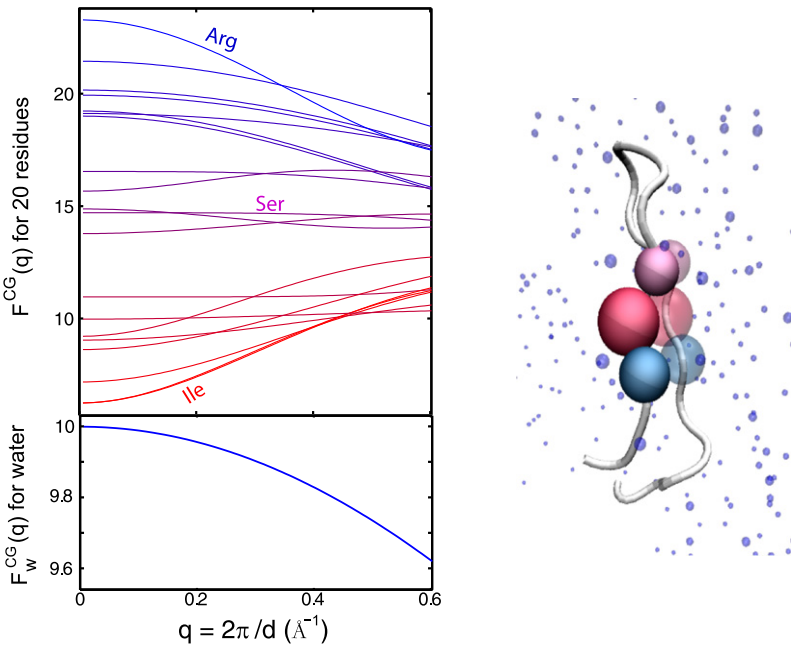

FIGURE 3 (*Top*) Effective residue-based scattering structure factors for 20 residues derived from a set of high-resolution crystal structures from the Protein Data Bank (PDB) (Eq. 5). (*Bottom*) The blue curve is the theoretical scattering factor of a TIP3P model water before the weighting (Eq. 6). An average scattering factor was calculated to account for backbone conformers and side-chain rotamers of each residue. For simplicity, we used an average over all residues in a set of high-resolution crystal structures. The data set consists of 434 protein chains derived from the PDB (as of July 2008). A large number of resulting atomic conformations were used for the averaging of each residue, ranging from 1308 for Cysteine, to 4400 for Proline, and to 8379 for Alanine. The ordering from large to small for 20 residues, according to the values of the intensity at $q \rightarrow 0$, is: Arg, His, Asp, Asn, Glu, Cys, Gln, Met, Trp, Tyr, Ser, Thr, Lys, Gly, Phe, Ala, Pro, Val, Leu, and Ile.

To model the electron density contrast in the primary hydration shell relative to the bulk phase, a weight is assigned for the scattering factor of dummy waters:

$$F^{\mathrm{CG}}(q) = w \times \mathrm{F}_{\mathrm{w}}^{\mathrm{CG}}(q). \tag{7}$$

As we shall see, the value of the weighting factor w is empirically calibrated using the experimental scattering data of lysozyme. With this strategy, the scattering from a given protein conformation is conveniently and accurately represented by its N residues via the Cα position and the surrounding M explicit water molecules via the Oxygen position. For this CG model of a protein with the accompanying hydration shell, the solution scattering can be calculated using the Debye formula

$$I^{\mathrm{CG}}(q) = \sum_{i,j=1}^{\mathrm{N+M}} F_i^{\mathrm{CG}}(q) F_j^{\mathrm{CG}}(q) \frac{\sin(qr_{ij})}{qr_{ij}}, \tag{8}$$

where $F_i^{\mathrm{CG}}(q)$ values are the effective CG scattering factors for both amino acids and water molecules (Eqs. 5 and 7). Thus, for a given protein conformation, a CG model of protein for scattering is achieved as a chain of N residues at their Cα positions and a layer of M dummy waters in the primary solvation shell.

Modeling protein flexibility in solution

The randomness of protein orientations in SAXS measurements requires spherical averaging in the theoretical framework. An additional level of disorder arises from the conformational flexibility of the protein. Proteins in solution fluctuate among accessible conformations, and the observed scattering reflects this ensemble. This is accomplished by the ensemble average of scattering $I(q) = \langle I^{\mathrm{CG}}(q) \rangle_{\mathrm{MD}}$, where $\langle \cdots \rangle_{\mathrm{MD}}$ stands for an MD average over the ensemble of structures around the local free energy minimum of folded proteins. Alternatively, they could be sampled by computational techniques such as Monte Carlo simulations, or normal mode analysis.

All aspects of the computation can be easily incorporated by sampling the local configurational space with CG simulations as a method of choice (37–42). In such a model, a protein is treated as a chain of Cα atoms with Lennard-Jones potentials to stabilize the native folded conformation. More can be found in Computational details, below.

Finally, an average scattering pattern of $I(q)$ is computed by taking into account 1), the effective scattering factors for amino acids and water molecules (with a proper weight); and 2), an ensemble of folded structures generated from the CG simulations that allow the protein to fluctuate around the native conformations.

Computational details

The protein conformational flexibility is modeled by using an ensemble of structures extracted from MD simulations with a simplified model built from the native conformations. The energy function for a Cα-based CG model is similar to the one used in Yang et al. (42), i.e.,

$$\begin{aligned} E = & \sum_{\text{bonds}} K_r (r - r_0)^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 \\ & + \sum_{\text{dihedrals}} K_\phi^{(n)} (1 - \cos(n(\phi - \phi_0))) \\ & + \sum_{\text{contacts}} \varepsilon_1 \left[5 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 6 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{10} \right] \\ & + \sum_{\text{repulsive}} \varepsilon_2 \left(\frac{\sigma_{\mathrm{o}}}{r_{ij}} \right)^{12}, \end{aligned} \tag{9}$$

where K_{r}, K_θ, and K_ϕ are the force constants of the bond, angle, and dihedral angle for adjacent C_α atoms, respectively, and we chose $K_{\mathrm{r}} = 100$ kcal/mol, $K_\theta = 20$ kcal/mol, $K_\phi^{(1)} = 1$ kcal/mol, and $K_\phi^{(3)} = 0.5$ kcal/mol. All native folded structures with r_0, θ_0, ϕ_0, and σ_{ij} are taken from the PDB. The value σ_{ij} is the distance of a pair of residues that are in contact in the native state. We chose $\varepsilon_1 = 1$ kcal/mol for native contacts and $\varepsilon_2 = 0.001$ kcal/mol for repulsive interactions for all pairs of residues that are not in contact in the native state ($\sigma_{\mathrm{o}} = 3.8$ Å).

This CG model is simulated by a Langevin dynamics with a leveraged friction coefficient and an increased time step. The friction coefficient was set to 50 ps^{-1} and a time step to 0.01 ps. The value of friction coefficient for Cα atoms was chosen to mimic the friction for the all-atomic-detailed residues (43). All simulations were carried out at a temperature of 300 K for a period of 1–5 ns.

For each configuration generated from CG simulations, a layer of water molecules is placed at the surface of the protein. This placement is done by removing water molecules overlapping with protein atoms from a large equilibrated TIP3P waterbox. The water molecules are inserted in a density of bulk solvent, $\rho_{\mathrm{s}} = 0.334\ e/\mathrm{Å}^3$. Finally, only those water molecules with positions 3.5–6.5 Å away from the protein Cα atoms are kept, to represent the hydration shell (Fig. 1 and Fig. 4).

RESULTS AND DISCUSSION

Here, we first derive knowledge-based coarse residue structure factors for all 20 residues and then calibrate the scattering of water molecules in the hydration shell using the well-studied protein lysozyme. We finally apply the CG method to several representative protein folds and to proteins with multiple biological conformational states.

Coarse-grained residue structure factors

Equation 5 was used to compute the effective residue-based scattering structure factors for 20 residues. The coarse-graining procedure was based on a set of high-resolution crystal structures. A set of 434 protein structures was derived from the PDB (as of July 2008) by using the PISCES program (35), based on the criteria: 1), sequence identity <10%; 2),

protein chain length from 40 to 10,000; 3), resolution <1.8 Å; and 4), *R*-factor value <0.15. This results in a large number of atomic conformations for each residue, ranging from 1308 for Cysteine, to 4400 for Proline, and to 8379 for Alanine. The scattering factor was derived from an average over all these conformers to account for different backbone and side-chain orientations of each residue (see Eq. 5).

Fig. 3 shows the CG residue scattering factors. The calculations were based on all appropriate conformers that are available in a subset of structure deposited in PDB using Eq. 5. The ordering of scattering intensity from large to small for 20 residues, according to the values at $q \rightarrow 0$, is: Arg, His, Asp, Asn, Glu, Cys, Gln, Met, Trp, Tyr, Ser, Thr, Lys, Gly, Phe, Ala, Pro, Val, Leu, and Ile. For instance, the residue of Arginine has a positive electron density relative to the bulk solvent, whereas Isoleucine has a relative negative density. Therefore, a CG scattering method can be constructed as a chain of effective residues at their Cα positions and having the effective structure factors derived.

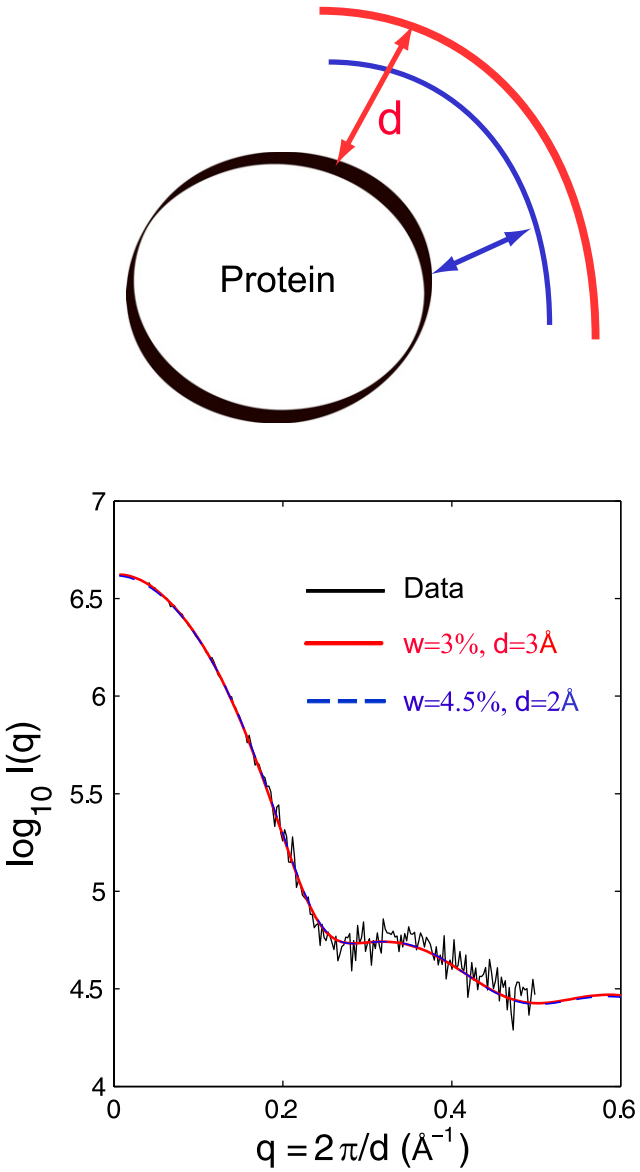

FIGURE 4 The choice of the thickness of the hydration layer. A combination of varying the thickness (d) of the layer and the weighting factor (w) of water molecules in the layer to maintain the overall net density yields very similar scattering results for lysozyme. Shown are two examples with $w = 3\%$, $d = 3$ Å (*red*) and $w = 4.5\%$, $d = 2$ Å (*blue*), respectively.

Lysozyme: a model system to calibrate the hydration shell

We used the HEW lysozyme (PDB code: 6LYZ) as our test case, since its SAXS data are publicly available through the CRYSOL package (9). We first simulated the lysozyme by a CG MD following Eq. 9 and then solvated the protein for each snapshot extracted from simulations by placing a layer of water molecules with an initial density of $\rho_s \sim 0.334$ e/Å^3 (see details above). Finally, the average scattering was calculated with CG residue scattering factors and a proper weight for dummy waters (to be determined) using Eq. 8.

Fig. 5 shows that the theoretical scattering intensity for the lysozyme with different lengths of simulation time from 1 ns to 5 ns. Clearly, the protein conformational flexibility is

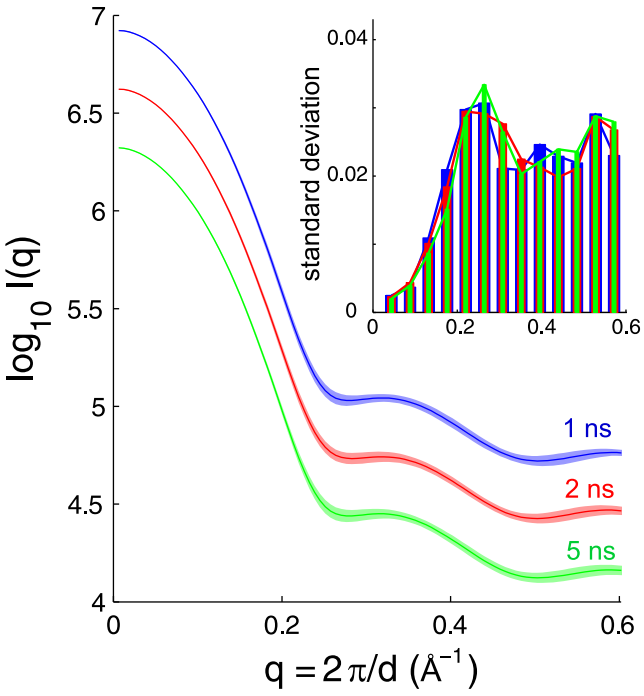

FIGURE 5 An ensemble average of configurations for scattering of the HEW lysozyme. The curves were calculated from an ensemble of snapshots taken from a simulation period of time of 1 ns, 2 ns, and 5 ns, respectively. The shades of each curve represent the standard deviations at each q position ($q = 2\pi/d$), which are enlarged by the plot in the inset. Comparison of these curves suggests that it is representative to take the scattering with an average over a period of 2-ns MD simulations. The weighting factor for dummy waters, $w = 3\%$ (Eq. 7), was used for the calculations. The intensity was plotted in a log-scale. For clarity, the curves with difference were shifted along the vertical direction.

reflected by the standard deviations (and the averages as well) of $I(q)$, which are represented by the shades of each curve (and the inset). In the low q region, where the overall protein shape is not sensitive to the conformation flexibility, the standard deviation of $I(q)$ is quite small. It gets larger in the higher q region, where the contributions of internal detailed fluctuations begin to dominate. For proteins with large conformational flexibility in solution, this larger uncertainty of $I(q)$, together with an intrinsic low signal/noise ratio in the high-q regions, may contribute to very noisy experimental observations as q increases.

Fig. 5 also shows that the standard deviations of $I(q)$ start to converge with a length of 2-ns simulations in the case of lysozyme. Although the convergence of the length of simulations has to be examined by a case-by-case basis for different proteins, we take as representative an ensemble average over a period of 2-ns simulations for the following discussion. In these calculations, the weighting factor for dummy waters $w = 3\%$ (Eq. 7) was used for the calculations, where the value of the factor is calibrated as follows.

To account for the local electron density difference in the hydration shell relative to the rest of bulk solvent, we modeled such a density contrast by assigning a proper weighting factor for the dummy water scattering according to Eq. 7. In this equation, there is one free parameter, w, that remains to be determined, to calculate the scattering curve. Fig. 6 shows the theoretical scattering intensities with several different weighting factors for waters of $w = 0\%$, 3%, and 10%. The weighting factor w for dummy waters reflects the excess electron density of the hydration shell relative to bulk solvent. In other words, the density of the shell is effectively w greater than that of the rest of bulk solvent. Fitted to the scattering data of lysozyme, a proper weight of $w = 3\%$ was chosen to reflect the relative difference in the primary solvent shell. Remarkably, the CG approach with a single fitting parameter for dummy waters can well reproduce the lysozyme data $I(q)$ up to $q = 0.5$ Å^{-1}. The high accuracy of reproducing the experimental scattering appears to be due to a combination of both the density difference and conformation flexibility.

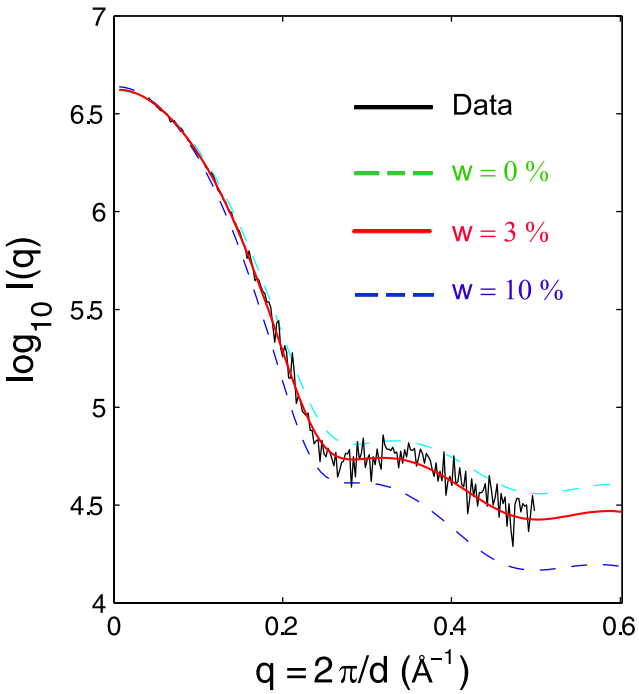

FIGURE 6 The weighting factor (w) for water molecules in the hydration layer with w of 0%, 3%, and 10%, respectively. The calculations were from 2-ns MD simulations. The theoretical curves were compared with the experimental data taken from the CRYSOL package. From the plot, a weight of $w = 3\%$ is chosen to fit the experimental curve. The data curve was shifted along the vertical direction to achieve an optimal overlap at the low q region.

We also note that different choices of the thickness of the hydration give rise to very similar scattering curves. Fig. 4 shows that a combination of varying the thickness (d) of the layer and the weighting factor (w) of water molecules in the layer to maintain the overall net density yields very similar scattering results for lysozyme, by two examples: with $w = 3\%$, $d = 3$ Å (*red*) and $w = 4.5\%$, $d = 2$ Å (*blue*), respectively. This would potentially further reduce the computational cost by using a thinner layer but with higher weighting factor for water molecules. From a physical consideration, we set the thickness of the hydration shell equal to 3 Å for the rest of discussion.

The CG residue scattering computational method can be advantageous. This CG calculation is much faster than the all-atom scattering calculations (9,11,12). For example, in the framework of the Debye formula, it is $\sim N^2$ times faster to compute the scattering of a protein without surrounding water molecules (N is the average number of atoms per residue). It should be noted that this advantage is less pronounced when explicit dummy water molecules are included in the calculation. Here, we make a very brief comparison of the CRYSOL calculations with our CG results. We computed the intensity of lysozyme from the widely used all-atom CRYSOL calculation, in which the atomic details for scattering factors were included. The CRYSOL calculation was carried out with default parameters. Fig. 7 shows the difference between the CG model and CRYSOL for lysozyme. Comparison shows that the CRYSOL calculation with the default parameters (solvation shell electron density 0.030 e/Å^3, i.e., 10% of bulk solvent electron density ρ_s) gives a very good intensity fit at the low q region, but shows a systematic shift from experiments at high q, whereas the CG model with $w = 3\%$ accurately reproduces the scattering curve. Putnam et al. also pointed out that this adjusted parameter with less density contrast (solvation shell electron density 0.010 e/Å^3, i.e., 3% of ρ_s) in CRYSOL gives a better scattering curve compared with data (1). This notion is also supported by our CG residue-based scattering calculations. Although there is room for refinement of the weight factor w when more reliable SAXS data from additional protein samples become available, we use $w = 3\%$ for the following discussion.

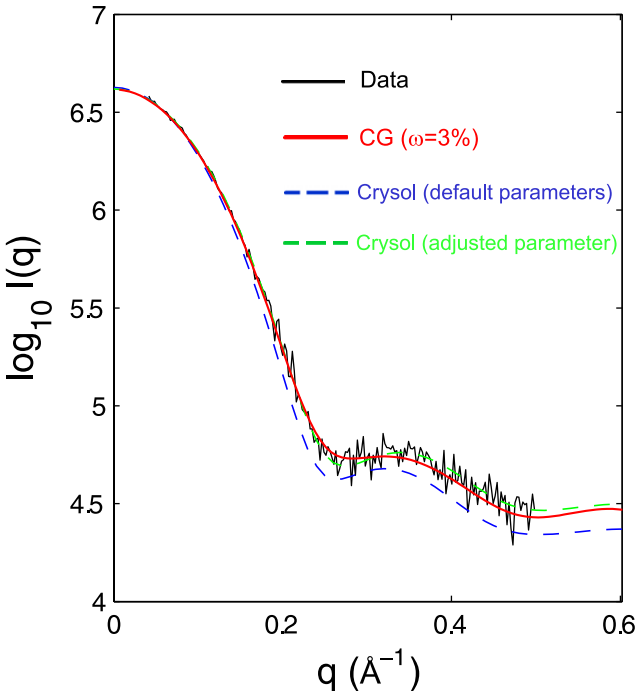

FIGURE 7 Theoretical scattering profiles of lysozyme from the CG model with a weight of $w = 3\%$ (*red*) and from CRYSOL (*blue* and *green*). The calculations from CRYSOL were performed with the default parameters (solvation shell contrast = 0.030 e/Å^3, 10% of bulk solvent electron density ρ_s) and an adjusted parameter (solvation shell contrast = 0.01 e/Å^3, 3% of ρ_s). The scattering curves were shifted along the vertical direction to achieve an optimal overlap at the low q region.

Comparison of our residue-based method with other simplified models is quite clear. Our method requires no additional computational cost to the simple Cα model (21,25,26), but includes effective residue structure factors and thus has a better accuracy (data not shown). In contrast to the DR model (23), our CG method requires no use of a correction function of q for scattering factors in order to reproduce the scattering patterns derived from the all-atom CRYSOL calculations. Instead, a knowledge-based, coarse-residue structure factor for each amino acid was calculated, and a single parameter for the density contrast was fitted for the solvent layer. Nonetheless, our CG method can well reproduce the experimental scattering of lysozyme, which provides us with a reasonable start point for further investigation.

It is worth noting that at high concentrations, interparticle correlations may lead to an interparticle form factor that is different from unity and will modulate the observed scattering at very small angles (44–46). In this article, since we assume that the scattering is from a dilute solution of protein particles, these interparticle effects are negligible.

Scattering characterization of protein folds

We now have a working model for calculating the scattering intensity from atomic models. Such a calculation can be very useful in many scenarios. For example, combined with SAXS data, the CG model can be used to model the biological assembled structures of multidomain complexes in cases where high-resolution crystal structures of each individual domain are known. Such an example will be presented in future communication. Currently, calculations from known atomic models can provide a scattering-signature of each protein fold. In fact, such an effort has been put forward by building a database of the CRYSOL-calculated scattering curves for a portion of structures deposited in the PDB (47,48). Similar efforts have been performed to characterize protein folds using wide-angle scattering calculations (49). In general, such a theoretical effort could be potentially useful for providing the ranking scores for experimental scattering data and further identifying top candidates from this kind of database for refinement. From this point of view, despite the possibility that multiple protein folds could share a similar scattering curve, we envision that the CG model can serve for a similar purpose to achieve scattering characterization of different protein folds.

Fig. 8 shows the scattering for several representative protein folds/structures, including α-helical, β-strand, and multidomain proteins. For a quantitative assessment of conformation flexibility, three scattering curves for each protein are computed and reported from 2-ns CG MD simulations. Conformational flexibility from simulations is reflected in the average (*lines*) and standard deviations (*shades*) in each curve. Three proteins are used to illustrate the scattering patterns of α-helical proteins.

In general, the SAXS scattering pattern contains less information than the pattern in crystallography, although it is still rich in details about the overall shape and internal structure of a macromolecule. In the low q region, the scattering can be used to measure the protein radius of gyration (R_G) by the Guinier approximation (50), $I(q) \propto e^{-q^2R_G^2/3}$. In the q region beyond R_G where the intensity starts to fall off, $I(q)$ shows a systematic trend for folded proteins, $I(q) \propto q^{-d}$, referred to as Porod's law (51,52). A value of $d = 4$ was found for many folded single-domain proteins. For multidomain complexes, our experience demonstrates that scattering from large domain-domain separations causes a modulation of the curve in this region. As q further increases, the power-law pattern breaks down where more detailed substructures start to contribute to the total scattering profile. These peaks are the collective contributions of scattering due to a spatial separation between large groups of atoms, such as the domain-domain separation and the secondary-structure packing.

Several interesting features might be noted from the theoretical patterns. First, the low-angle scattering contains information about protein size and overall shape/envelope: the larger the protein, the greater the scattering intensity at $q \rightarrow 0$ (in theory, $I(q = 0)$ is proportional to protein size (1–3)). For example, Bcl-X with a total of 196 residues

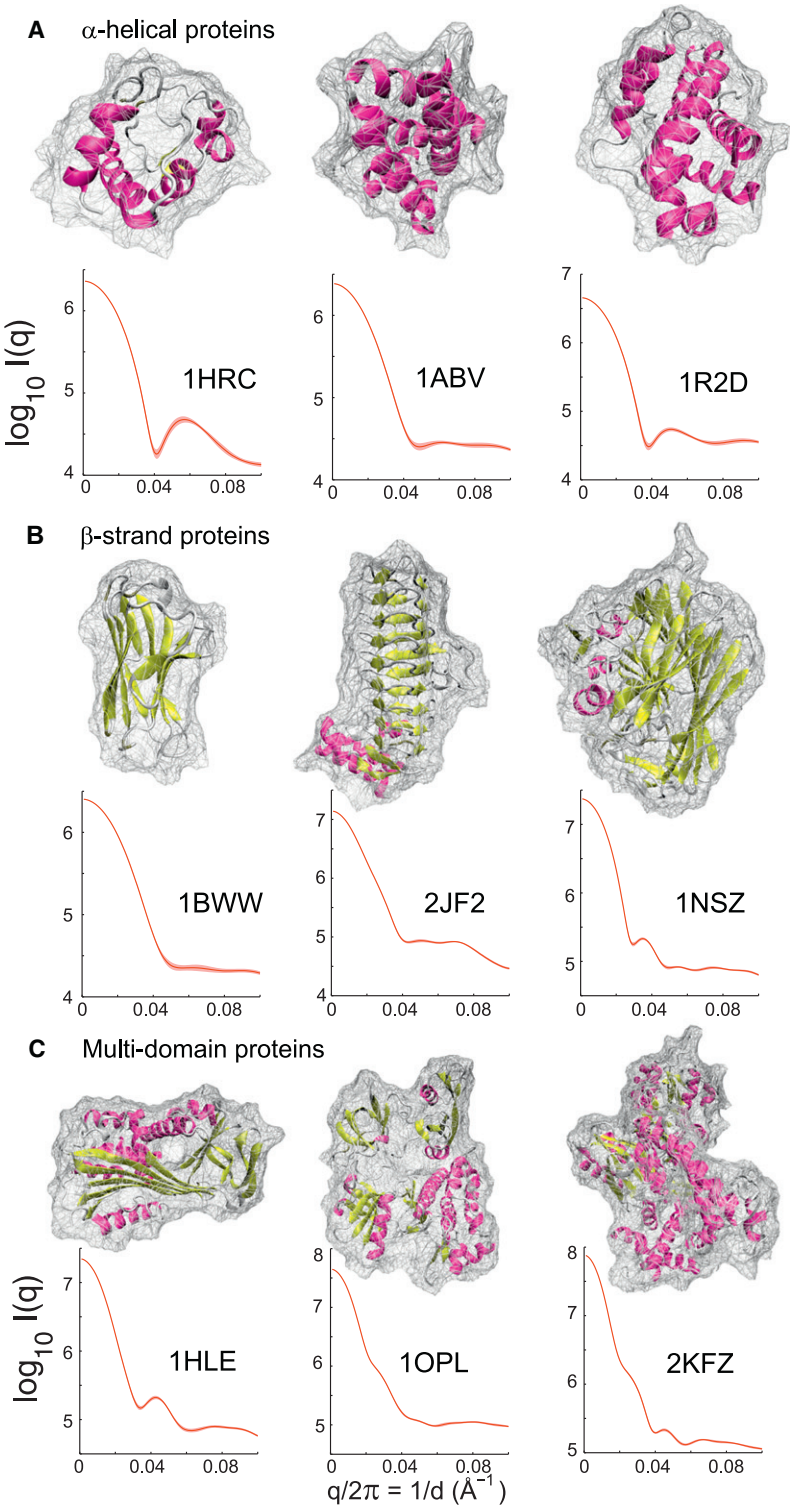

FIGURE 8 Scattering characterization of protein folds: α-helical, β-strand, and multidomain proteins. The low-angle scattering contains information about protein size and overall shape; the wider-angle scattering provides information about secondary-structure packing and domain motions. (*A*) The α-helical proteins: cytochrome *c* (PDB code: 1HRC (70)), ATPsynthase (PDB code: 1ABV (71)), and Bcl-X (PDB code: 1R2D (72)). (*B*) The β-strand proteins: immunoglobulin (PDB code: 1BWW (73)), acyltransferase (PDB code: 2JF2 (74)), and galactose mutarotase (PDB code: 1NSZ (75)). (*C*) Multidomain proteins: serpin (PDB code: 1HLE (76)), c-Abl (PDB code: 1OPL (59)), and DNA polymerase (PDB code: 2KFZ (60)). Each curve was calculated from an ensemble of snapshots taken from a 2-ns MD simulation trajectory. A log-scale of the scattering intensity was used for all the proteins.

(PDB code: 1R2D) has a higher scattering intensity than cytochrome *c* (104 residues).

Second, the higher angle scattering provides information about secondary-structure packing, e.g., the helix-helix organization in the case of all-α proteins as observed as in cytochrome *c* and Bcl-X. However, such a peak is not generic for the proteins in the family of α-helical proteins. For example, the curve is relatively flat in ATPsynthase (PDB code: 1ABV) or a peak is found at a lower angle at $\sim q/2\pi = 1/d \sim 0.05$ Å^{-1} in Bcl-X. Further wider-angle scattering ($q > 0.6$ Å^{-1}) was not investigated here, but has been studied by several experimental groups (53,54).

Similar features are observed in β-strand proteins including the all-β immunoglobulin (PDB code: 1BWW), the β-helix acyltransferase (PDB code: 2JF2), and the supersandwich fold of galactose mutarotase (PDB code: 1NSZ). In particular, the β-helix fold has recently become a very interesting research focus, in part because it has been proposed as a structural candidate for amyloid proteins such as the prion protein (55–57) and the Aβ protein (58). It suggests that SAXS combined with computation has a significant potential of elucidating the basic structural details of cross-β fingerprints implicated in many amyloid diseases such as Alzheimer's. In the case of acyltransferase, the scattering curve displays a plateaulike flat pattern in a quite wide q-range ($1/d \sim 0.04$–0.07 Å^{-1}), before falling off at higher q ($1/d \sim 0.08$ Å^{-1}). It differs from the flat pattern in the all-β immunoglobulin, where the scattering intensity does not start to fall off even at $1/d \sim 0.09$ Å^{-1}.

SAXS scattering appears to offer significant potential advantage for examining the structures of multidomain proteins. Shown in Fig. 8 are the theoretical scattering patterns from serpin (PDB code: 1HLE), c-Abl (PDB code: 1OPL), and DNA polymerase (PDB code: 2KFZ). As mentioned earlier, rich scattering information such as one or multiple peaks at the power-law or high q regions can be observed because of the collective separation of two large groups. This is clearly represented in all three cases because of the domain-domain organization. For example, there is a peak at $\sim 1/d \sim 0.04$ Å^{-1} in serpin, which represents a major separation of $d \sim 25$ Å between two domains. We note that the c-Abl tyrosine kinase has been the target of drug design for cancer treatment (59). The scattering of c-Abl shows a very detailed pattern in a wide q-range (e.g., $1/d \sim 0.03$ Å^{-1} and $1/d \sim 0.07$ Å^{-1}), reflecting the complex assembly among two regulatory domains and one catalytic domain. Similarly, peaks are superimposed on the power-law region and the high q region for a larger complex of the DNA polymerase (60). Thus, combined with computation including MD simulations, the solution scattering can have a great potential in understanding how multiple domains assemble in physiological conditions.

Use of SAXS for fold recognition

As described above, different proteins display distinct scattering patterns by protein size at $q \rightarrow 0$, R_G at low q, structural packing at higher q, etc. Such distinctions suggest that a SAXS scattering curve serve as a characteristic or semi-signature of each protein fold/structure. As mentioned earlier, the knowledge about the theoretical scattering could be potentially useful by creating a database of theoretical scattering curves, similar to the CRYSOL-based DARA (47). Ideally, experimental scattering from an unknown fold is fitted to the precalculated theoretical scattering curves to obtain a list of top hits by ranking. A ranking score for such a measure could be developed based on a χ^2 parameter

$$\chi^2 = \sum_{i_q=1}^{N_q} \frac{\left(\log I^{\exp}(q) - \log I^{CG}(q) - \Delta\right)^2}{\sigma^2(q)}, \qquad (10)$$

where N_q is the number of data points in the scattering curve and $N_q = 100$ was used for theoretical calculations throughout the rest of the article. The value Δ is a normalization factor, which is used to offset the difference of scattering intensity at $q \rightarrow 0$. The value $\sigma(q)$ is experimental uncertainty of log $I^{\exp}(q)$ or simulated standard deviation of log $I^{CG}(q)$, in cases where the experimental errors are not available.

To illustrate the concept of fold recognition by the use of SAXS, we first computed the pairwise χ^2 derivations between the scattering signature of the nine protein folds above-described. Table 1 shows that the χ^2 parameter ranges from 10^3 to 10^6, suggesting that there is a large separation between different protein folds. Such a large separation makes possible the construction of a theoretical database for fold recognition. To illustrate this concept, we used two representative proteins, the Bcl-2 homolog from myxoma virus (PDB code: 2O42) and the YDCK from *Salmonella cholerae* (PDB code: 2F9C), which have similar structures to Bcl-X (PDB code: 1R2D) and acyltransferase (PDB code: 2JF2), respectively. The computed scattering from these two proteins were ranked against the database of the

TABLE 1 The pairwise χ^2 distances in the scattering space between the nine proteins as shown in Fig. 8

	α-Helical proteins			β-Strand proteins			Multidomain proteins		
$\chi^2(\times 10^6)$	1HRC	1ABV	1R2D	1BWW	2JF2	1NSZ	1HLE	1OPL	2KFZ
1HRC	0	0.0031	0.0110	0.0038	0.1037	0.1574	0.2038	0.4043	0.6624
1ABV		0	0.0098	0.0016	0.0752	0.1275	0.1510	0.2882	0.4609
1R2D			0	0.0051	0.0382	0.0537	0.0770	0.1850	0.3291
1BWW				0	0.1759	0.2877	0.3464	0.6307	1.0118
2JF2					0	0.0581	0.0452	0.1075	0.2069
1NSZ						0	0.0149	0.1217	0.3385
1HLE							0	0.0273	0.0811
1OPL								0	0.0119
2KFZ									0

The large separation between them suggests that a database of scattering would be useful for fold recognition. The calculations were based on Eq. 10, where the standard deviation of theoretical curves was used for $\sigma(q)$.

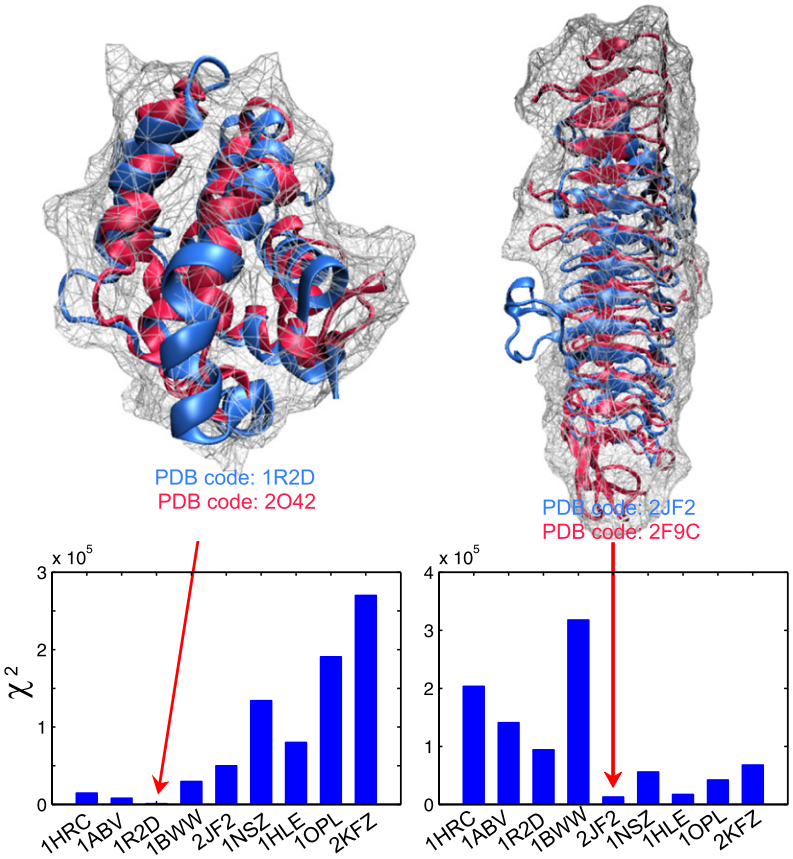

FIGURE 9 Two illustrative examples for fold recognition by χ^2 parameters. (*Left*) The computed scattering curve from the Bcl-2 homolog from myxoma virus (*red*; PDB code: 2O42 (77)) best-matches that of Bcl-X (*blue*; PDB code: 1R2D). (*Right*) The computed scattering curve from the YDCK from *Salmonella cholerae* (*red*; PDB code: 2F9C) best-fits with that of acyltransferase (*blue*; PDB code: 2JF2); both have a similar β-helical fold. In both cases, the sequence identities are quite low, 10.7% and 16.9%, respectively, according to the CE alignment calculations (61). The calculations for χ^2 were based on Eq. 10 and the standard deviation of scattering curves was used for $\sigma(q)$.

precomputed nine curves, according to χ^2 (Eq. 10). Fig. 9 shows that the best hit for both cases is indeed the one which is the best match (marked by *arrows*), according to the CE alignment calculations (61). We note that in both cases proteins are similar in fold but very different in sequence, e.g., the sequence identities are 10.7% and 16.9%, respectively. These encouraging results suggest that SAXS might provide an alternative approach for protein structure prediction by taking advantage of the ease of use of solution scattering to support and complement current homology modeling or ab initio protein structure predictions (62). We envision that the concept of best-fitting to a theoretical scattering database of all known folds could play an important role in fold recognition. For this purpose, the rapid determination of scattering profiles from our CG method can provide a fast and efficient way to create a database of SAXS profiles for all appropriate protein folds as deposited in the PDB.

Further application of SAXS data to the so-called natively disordered proteins has been shown to elucidate their structural features (e.g., (63,64).). In this case, because a large ensemble of unfolded structures must be sampled, our rapid CG computational method might provide an efficient tool for characterizing structural features of disordered states. Incorporation of SAXS data with NMR data for structure refinements (65,66) is beyond the scope of this article, but our CG method can provide an alternative approach for scattering calculations, especially for low- and medium-angle scattering.

Scattering characterization for multiple conformational states

Protein can adopt multiple conformational states in equilibrium under specific physiological conditions. Specific binding to ligands, substrates, or target proteins can shift the population from one state to another (19). In cases where each individual state is well defined, theoretical scattering can be used to characterize the conformational states from atomic protein models. Such a theoretical scattering calculation can be an important tool to deconvolute the relative population of each state in the SAXS sample with mixing states in solution. Here, several examples are given for theoretical scattering patterns of distinct conformational states from given atomic models.

Fig. 10 shows the theoretical scattering curves of two distinct conformational states from three proteins: transferrin, calmodulin, and ParM. In transferrin, ligand-induced conformational change occurs between two lobes when iron binds. The structures of the apo- and holoforms are shown in blue and red, respectively. The differences in theoretical curves are at $1/d \sim 0.03$ Å^{-1} and $1/d \sim 0.06$–0.08 Å^{-1}, which suggests a large-scale conformational change upon the binding.

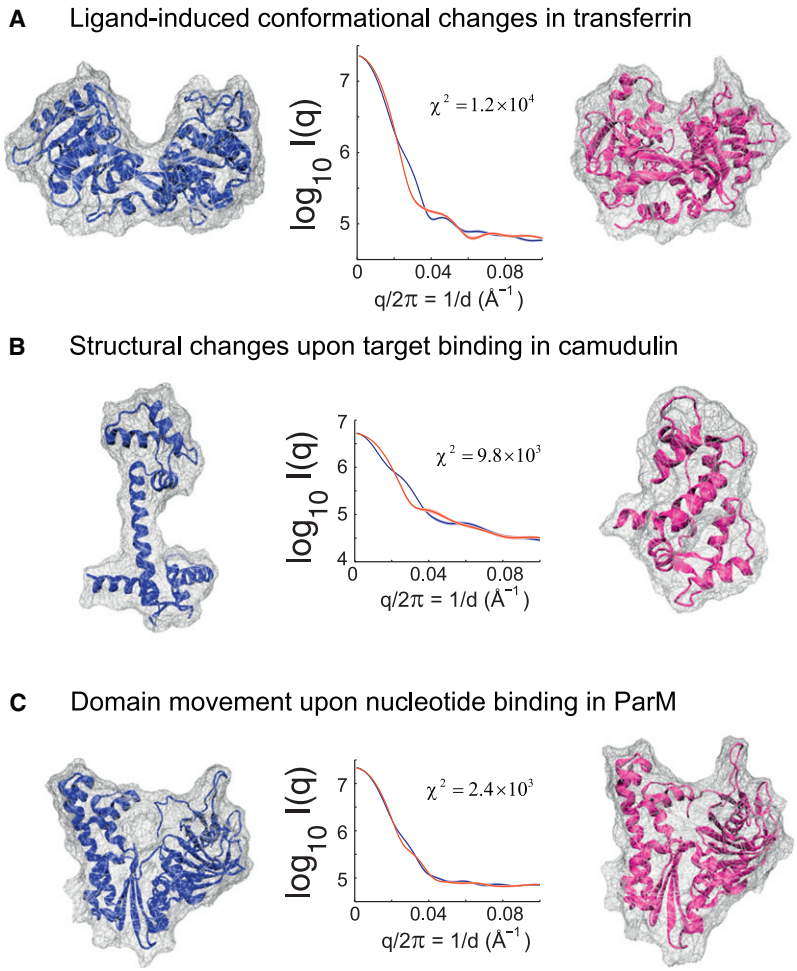

FIGURE 10 Scattering difference of multiple conformation states. The differences in scattering curves are found in a wide q-range from low to high, suggesting that SAXS can be used to investigate the large-scale conformational motions in solution. The differences are also reflected by a χ^2 parameter between two curves. (*A*) Ligand-induced conformational change in transferrin before (*blue*) and after (*red*) the ion binding (PDB codes: 1BP5 (78) and 1A8E (79), respectively). (*B*) The dumbbell shape (*blue*) and the compact shape (*red*) of calmodulin (PDB codes: 1CLL (67) and 1CDL (80), respectively). (*C*) Domain-domain conformations before and after nucleotide binding in ParM (PDB codes: 1MWK (81) and 1MWM (81), respectively). All scattering curves in a log-scale were calculated from an ensemble of snapshots taken from an MD simulation period of time of 2 ns. The χ^2 parameter was calculated using Eq. 10, where the standard deviation of the theoretical scattering of the left state was used for $\sigma(q)$.

Another well-studied example is the Ca^{2+}-bonded calmodulin where a major conformational change occurs when it binds to a target protein (67). The difference in theoretical curves at low q ($1/d \sim 0.01$–0.02 Å^{-1} and $1/d \sim 0.03$ Å^{-1}) in Fig. 10 clearly represents the distinct protein shapes (e.g., R_G) and helix-helix arrangements. This arises from the range of forms calmodulin displays in solution, from extended to more compact. The equilibrium can be shifted from one to another, dependent on whether it is in solution, a crystal lattice, or binding to specific agents. Thus, SAXS has been a method of choice for studying such a multistate calmodulin in both experimental and computational aspects (68,69).

A similar feature is also found in theoretical scattering curves in the apo- and holoforms of ParM, a member of the actin filament protein family, where domains move upon nucleotide binding. Scattering differences can be found in several places such as $1/d \sim 0.03$ Å^{-1} and $1/d \sim 0.06$ Å^{-1} (Fig. 10). Again, this demonstrates the usefulness of SAXS for characterizing distinct functional states in solution. The χ^2 calculations also indicate scattering differences in distinct states (Fig. 10).

To summarize, several multiconformational-state proteins examined here show that a major difference in scattering of distinct states is observed in or near the power-law region of scattering curves. Such distinct patterns can be detected by SAXS experiments. This suggests that the solution scattering at relative small angles has the capability of identifying the assembly mechanisms of multiple-state proteins, often with multiple domains. These kinds of SAXS experiments are obviously attractive because of the difficulties of growing crystals for large protein complexes.

CONCLUSION

SAXS is an increasingly important technique for characterizing macromolecular folds, conformations and assembly states in physiological conditions. It can provide low-resolution structural information without the challenges and limitations of crystallography or solution NMR. In principle, the scattering data can be used as an input for computational modeling to reconstruct the structures for multiprotein complexes and to deconvolute the equilibrium population of each conformational state of proteins in solution. However, a reliable and efficient computational approach is needed to achieve this goal.

A theoretical CG model was developed to compute the scattering pattern from a protein in a given conformation.

The model is residue-based, but the scattering for each amino acid was built from atomically detailed conformers. Such a CG representation for calculating scattering is advantageous, because it significantly reduces the computational cost and it can be combined with CG simulations that can be used to sample broad configurational spaces exhibited by many large complexes. The CG representation of the protein takes advantage of the low-resolution character of SAXS. The computational methods were further illustrated by characterizing a variety of protein folds and multiple conformational states. Preliminary tests show that a given fold can be detected via the scattering signature of the protein. This suggests that the structural information from SAXS, when combined with computations, could provide a powerful route for rapid fold recognition and shape reconstruction of large macromolecular complexes.

The program for this rapid coarse residue-based computational method for proteins will be released under the GNU General Public License with the code name of Fast-SAXS.

APPENDIX

In this Appendix, the procedure of simplifying each amino acid into a "glob" in Eq. 5 is demonstrated in the case of a group of spherical atoms. The scattering from these atoms within a given protein conformation is given by Eq. 4, which can be rewritten as

$$I(q) = \sum_{g=1}^{N_G} F^2(q) + \sum_j \sum_{j'} f_j'(q) f_{j'}'(q) \frac{\sin\left(qr_{jj'}\right)}{qr_{jj'}}, \quad (11)$$

where N_G is the number of amino acids and $F(q)$ is the scattering factor of the g^{th} residue in the protein. The notations j and j' refer to atoms in different residues, and $r_{jj'}$ are the distances between atoms j and j'.

In the case of spherical atoms where scattering factors are independent of direction, Harker pointed out that Eq. 11 can be essentially given by (27)

$$I(q) = \sum_{g=1}^{N_G} F^2(q) + \sum_g \sum_{g' \neq g} F_g(q) F_{g'}(q) \frac{\sin\left(qr_{gg'}\right)}{qr_{gg'}}, \quad (12)$$

where $r_{gg'}$ values are the distances between amino acids g and g'.

We thank Jan Lipfert for very helpful comments and suggestions on the article; we also thank Albert Lau, Nilesh Banavali, Jaydeep Bardhan, and Franci Merzel for valuable discussions.

This work was supported by the National Institute of Health through grant No. CA-093577 and by a joint grant from the University of Chicago Cancer Center and Argonne National Laboratory (grant No. UCCC/ANL).

REFERENCES

1. Putnam, C. D., M. Hammel, G. L. Hura, and J. A. Tainer. 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.
2. Lipfert, J., and S. Doniach. 2007. Small-angle x-ray scattering from RNA, proteins, and protein complexes. *Annu. Rev. Biophys. Biomol. Struct.* 36:307–327.
3. Koch, M. H. J., P. Vachette, and D. I. Svergun. 2003. Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Q. Rev. Biophys.* 36:147–227.
4. Chu, B., and B. Hsiao. 2001. Small-angle x-ray scattering of polymers. *Chem. Rev.* 101:1727–1762.
5. Doniach, S. 2001. Changes in biomolecular conformation seen by small angle x-ray scattering. *Chem. Rev.* 101:1763–1778.
6. Perkins, S. J. 1988. Structural studies of proteins by high-flux x-ray and neutron solution scattering. *Biochem. J.* 254:313–327.
7. Forster, F., B. Webb, K. A. Krukenberg, H. Tsuruta, D. A. Agard, et al. 2008. Integration of small-angle x-ray scattering data into structural modeling of proteins and their assemblies. *j. mol. biol.* 382:1089–1106.
8. Bernado, P., Y. Pérez, D. I. Svergun, and M. Pons. 2008. Structural characterization of the active and inactive states of Src kinase in solution by small-angle x-ray scattering. *J. Mol. Biol.* 376:492–505.
9. Svergun, D., C. Barberato, and M. H. J. Koch. 1995. CRYSOL—a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystal.* 28:768–773.
10. Merzel, F., and J. C. Smith. 2002. SASSIM: a method for calculating small-angle x-ray and neutron scattering and the associated molecular envelope from explicit-atom models of solvated proteins. *Acta Crystallogr. D Biol. Crystallogr.* 58:242–249.
11. Tiede, D., R. Zhang, and S. Seifert. 2002. Protein conformations explored by difference high-angle solution x-ray scattering: oxidation state and temperature dependent changes in cytochrome *c*. *Biochemistry.* 41:6605–6614.
12. Tjioe, E., and W. T. Heller. 2007. ORNL_SAS: software for calculation of small-angle scattering intensities of proteins and protein complexes. *J. Appl. Cryst.* 40:782–785.
13. Petoukhov, M., and D. Svergun. 2007. Analysis of x-ray and neutron scattering from biomacromolecular solutions. *Curr. Opin. Struct. Biol.* 17:562–571.
14. Hubbard, S., K. Hodgson, and S. Doniach. 1988. Small-angle x-ray scattering investigation of the solution structure of troponin C. *J. Biol. Chem.* 263:4151–4158.
15. Svergun, D. I., S. Richard, M. H. J. Koch, Z. Sayers, S. Kuprin, et al. 1998. Protein hydration in solution: experimental observation by x-ray and neutron scattering. *Proc. Natl. Acad. Sci. USA.* 95:2267–2272.
16. Merzel, F., and J. C. Smith. 2002. Is the first hydration shell of lysozyme of higher density than bulk water? *Proc. Natl. Acad. Sci. USA.* 99:5378–5383.
17. Koizumi, M., H. Hirai, T. Onai, K. Inoue, and M. Hirai. 2007. Collapse of the hydration shell of a protein prior to thermal unfolding. *J. Appl. Cryst.* 40:s175–s178.
18. Boehr, D. D., D. McElheny, H. J. Dyson, and P. E. Wright. 2006. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science.* 313:1638–1642.
19. Vendruscolo, M., and C. M. Dobson. 2006. Dynamic visions of enzymatic reactions. *Science.* 313:1586–1587.
20. Makowski, L., D. J. Rodi, S. Mandava, D. D. Minh, D. B. Gore, et al. 2008. Molecular crowding inhibits intramolecular breathing motions in proteins. *J. Mol. Biol.* 375:529–546.
21. Walther, D., F. E. Cohen, and S. Doniach. 2000. Reconstruction of low-resolution three-dimensional density maps from one-dimensional small-angle x-ray solution scattering data for biomolecules. *J. Appl. Cryst.* 33:350–363.
22. Guo, D. Y., R. H. Blessing, D. A. Langs, and G. D. Smith. 1999. On "globbicity" of low-resolution protein structures. *Acta Crystallogr. D Biol. Crystallogr.* 55:230–237.
23. Svergun, D. I., M. V. Petoukhov, and M. H. J. Koch. 2001. Determination of domain structure of proteins from x-ray solution scattering. *Biophys. J.* 80:2946–2953.
24. Chacón, P., F. Morán, J. Díaz, E. Pantos, and J. Andreu. 1998. Low-resolution structures of proteins in solution retrieved from x-ray scattering with a genetic algorithm. *Biophys. J.* 74:2760–2775.

25. Zheng, W., and S. Doniach. 2002. Protein structure prediction constrained by solution x-ray scattering data and structural homology identification. *J. Mol. Biol.* 316:173–187.

26. Wu, Y., X. Tian, M. Lu, M. Chen, Q. Wang, et al. 2005. Folding of small helical proteins assisted by small-angle x-ray scattering profiles. *Structure.* 13:1587–1597.

27. Harker, D. 1953. The meaning of the average of $|F|^2$ for large values of the interplanar spacing. *Acta Crystallogr.* 6:731–736.

28. Guo, D. Y., G. D. Smith, J. F. Griffin, and D. A. Langs. 1995. Use of globic scattering factors for protein structures at low resolution. *Acta Crystallogr. A.* 51:945–947.

29. Bragg, L., and M. F. Perutz. 1952. The structure of hemoglobin. *Proc. Roy. Soc. A (Lond.).* 213:425–435.

30. Fraser, R. D. B., T. P. MacRae, and E. Suzuki. 1978. An improved method for calculating the contribution of solvent to the x-ray diffraction pattern of biological molecules. *J. Appl. Cryst.* 11:693–694.

31. Lee, S., and D. Eisenberg. 2003. Seeded conversion of recombinant prion protein to a disulfide-bonded oligomer by a reduction-oxidation process. *Nat. Struct. Biol.* 10:725–730.

32. Cromer, D. T., and J. B. Mann. 1968. X-ray scattering factors computed from numerical Hartree-Fock wave functions. *Acta Cryst. A.* 24:0567–7394.

33. Lau, A. Y., and B. Roux. 2007. The free energy landscapes governing conformational changes in a glutamate receptor ligand-binding domain. *Structure.* 15:1203–1214.

34. Debye, P. 1915. Dispersion of Roentgen rays. *Ann. Phys. (Leipzig).* 46:809–823.

35. Wang, G., J. Dunbrack, and L. Roland. 2003. PISCES: a protein sequence culling server. *Bioinformatics.* 19:1589–1591.

36. Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.

37. Nymeyer, H., A. E. García, and J. N. Onuchic. 1998. Folding funnels and frustration in off-lattice minimalist models. *Proc. Natl. Acad. Sci. USA.* 95:5921–5928.

38. Clementi, C., H. Nymeyer, and J. N. Onuchic. 2000. Topological and energetic factors: what determines the structural details of the transition state ensemble and "on-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* 298:937–953.

39. Koga, N., and S. Takada. 2001. Roles of native topology and chain-length scaling in protein folding: a simulation study with a Gō-like model. *J. Mol. Biol.* 313:171–180.

40. Cheung, M. S., A. E. Garcia, and J. N. Onuchic. 2002. Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Natl. Acad. Sci. USA.* 99:685–690.

41. Karanicolas, J., and C. L. Brooks. 2002. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.* 11:2351–2361.

42. Yang, S., J. N. Onuchic, and H. Levine. 2006. Effective stochastic dynamics on a protein folding energy landscape. *J. Chem. Phys.* 125:054910.

43. Yang, S., and B. Roux. 2008. Src kinase conformational activation: thermodynamics, pathways, and mechanisms. *PLoS Comput. Biol.* 4:e1000047.

44. Stradner, A., H. Sedgwick, F. Cardinaux, W. C. K. Poon, S. U. Egelhaaf, et al. 2004. Equilibrium cluster formation in concentrated protein solutions and colloids. *Nature.* 432:492–495.

45. Shukla, A., E. Mylonas, E. Di Cola, S. Finet, P. Timmins, et al. 2008. Absence of equilibrium cluster phase in concentrated lysozyme solutions. *Proc. Natl. Acad. Sci. USA.* 105:5075–5080.

46. Trewhella, J. 2008. The different views from small angles. *Proc. Natl. Acad. Sci. USA.* 105:4967–4968.

47. Sokolova, A., V. Volkov, and D. Svergun. 2003. Database for rapid protein classification based on small-angle x-ray scattering data. *Crystallogr. Rep.* 48:959–965.

48. Sokolova, A. V., V. V. Volkov, and D. I. Svergun. 2003. Prototype of a database for rapid protein classification based on solution scattering data. *J. Appl. Cryst.* 36:865–868.

49. Makowski, L., D. J. Rodi, S. Mandava, S. Devarapalli, and R. F. Fischetti. 2008. Characterization of protein fold using wide-angle x-ray solution scattering. *J. Mol. Biol.* 383:731–744.

50. Guinier, A., and G. Fournet. 1955. Small-Angle Scattering of X-Rays. Wiley, New York.

51. Glatter, O. 1977. A new method for the evaluation of small-angle scattering data. *J. Appl. Cryst.* 10:415–421.

52. Roe, R.-J. 2000. Methods of X-Ray and Neutron Scattering in Polymer Science. Oxford University Press, New York.

53. Hirai, M., H. Iwase, T. Hayakawa, K. Miura, and K. Inoue. 2002. Structural hierarchy of several proteins observed by wide-angle solution scattering. *J. Synchrotron Radiat.* 9:202–205.

54. Fischetti, R. F., D. J. Rodi, D. B. Gore, and L. Makowski. 2004. Wide-angle x-ray solution scattering as a probe of ligand-induced conformational changes in proteins. *Chem. Biol.* 11:1431–1443.

55. Govaerts, C., H. Wille, S. B. Prusiner, and F. E. Cohen. 2004. Evidence for assembly of prions with left-handed β-helices into trimers. *Proc. Natl. Acad. Sci. USA.* 101:8342–8347.

56. Yang, S., H. Levine, J. N. Onuchic, and D. L. Cox. 2005. Structure of infectious prions: stabilization by domain swapping. *FASEB J.* 19:1778–1782.

57. Kunes, K. C., S. C. Clark, D. L. Cox, and R. R. Singh. 2008. Left-handed β-helix models for mammalian prion fibrils. *Prion.* 2:81–90.

58. Guo, J. T., R. Wetzel, and Y. Wu. 2004. Molecular modeling of the core of Aβ amyloid fibrils. *Proteins.* 57:357–364.

59. Nagar, B., O. Hantschel, M. A. Young, K. Scheffzek, D. Veach, et al. 2003. Structural basis for the autoinhibition of c-Abl tyrosine kinase. *Cell.* 112:859–871.

60. Brautigan, C., S. Sun, J. Piccirilli, and T. Steitz. 1999. Structures of normal single-stranded DNA and deoxyribo-3′-S-phosphorothiolates bound to the 3′-5′ exonucleolytic active site of DNA polymerase I from *Escherichia coli*. *Biochemistry.* 38:696–704.

61. Shindyalov, I., and P. Bourne. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* 11:739–747.

62. Baker, D., and A. Sali. 2001. Protein structure prediction and structural genomics. *Science.* 294:93–96.

63. Zagrovic, B., J. Lipfert, E. J. Sorin, I. S. Millett, W. F. van Gunsteren, et al. 2005. Unusual compactness of a polyproline type II structure. *Proc. Natl. Acad. Sci. USA.* 102:11698–11703.

64. Wells, M., H. Tidow, T. J. Rutherford, P. Markwick, M. R. Jensen, et al. 2008. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc. Natl. Acad. Sci. USA.* 105:5762–5767.

65. Grishaev, A., J. Wu, J. Trewhella, and A. Bax. 2005. Refinement of multidomain protein structures by combination of solution small-angle x-ray scattering and NMR data. *J. Am. Chem. Soc.* 127:16621–16628.

66. Grishaev, A., V. Tugarinov, L. E. Kay, J. Trewhella, and A. Bax. 2008. Refined solution structure of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints. *J. Biomol. NMR.* 40:95–106.

67. Chattopadhyay, R., W. E. Meador, A. R. Means, and F. A. Quiocho. 1992. Calmodulin structure refined at 1.7 Å resolution. *J. Mol. Biol.* 228:1177–1192.

68. Heidorn, D. B., and J. Trewhella. 1988. Comparison of the crystal and solution structures of calmodulin and troponin C. *Biochemistry.* 27:909–915.

69. Vigil, D., S. C. Gallagher, J. Trewhella, and A. E. Garcia. 2001. Functional dynamics of the hydrophobic cleft in the N-domain of calmodulin. *Biophys. J.* 80:2082–2092.

70. Bushnell, G. W., G. V. Louie, and G. D. Brayer. 1990. High-resolution three-dimensional structure of horse heart cytochrome *c*. *J. Mol. Biol.* 214:585–595.

71. Wilkens, S., S. D. Dunn, J. Chandler, F. W. Dahlquist, and R. A. Capaldi. 1997. Solution structure of the terminal domain of the δ-subunit of the *E. coli* ATPsynthase. *Nat. Struct. Biol.* 4:198–201.

72. Manion, M. K., J. W. O'Neill, C. D. Giedt, K. M. Kim, K. Y. Z. Zhang, et al. 2004. Bcl-XL mutations suppress cellular sensitivity to antimycin A. *J. Biol. Chem.* 279:2159–2165.

73. Usón, I., E. Pohl, T. R. Schneider, Z. Dauter, A. Schmidt, et al. 1999. 1.7 Å structure of the stabilized REI$_y$ mutant T39K. Application of local NCS restraints. *Acta Crystallogr. D Biol. Crystallogr.* 55:1158–1167.

74. Ulaganathan, V., L. Buetow, and W. Hunter. Nucleotide substrate recognition by UDP-*n*-acetylglucosamine acyltransferase (LPXA) in the first step of lipid A biosynthesis. Accepted.

75. Thoden, J. B., J. Kim, F. M. Raushel, and H. M. Holden. 2003. The catalytic mechanism of galactose mutarotase. *Protein Sci.* 12:1051–1059.

76. Baumann, U., W. Bode, R. Huber, J. Travis, and J. Potempa. 1992. Crystal structure of cleaved equine leukocyte elastase inhibitor determined at 1.95Å resolution. *J. Mol. Biol.* 226:1207–1218.

77. Douglas, A. E., K. D. Corbett, J. M. Berger, G. McFadden, and T. M. Handel. 2007. Structure of M11L: a myxoma virus structural homolog of the apoptosis inhibitor, Bcl-2. *Protein Sci.* 16:695–703.

78. Jeffrey, P., M. Bewley, R. MacGillivray, A. Mason, R. Woodworth, et al. 1998. Ligand-induced conformational change in transferrins: crystal structure of the open form of the N-terminal half-molecule of human transferrin. *Biochemistry.* 37:13978–13986.

79. MacGillivray, R., S. Moore, J. Chen, B. Anderson, H. Baker, et al. 1998. Two high-resolution crystal structures of the recombinant N-lobe of human transferrin reveal a structural change implicated in iron release. *Biochemistry.* 37:7919–7928.

80. Meador, W., A. Means, and F. Quiocho. 1992. Target enzyme recognition by calmodulin: 2.4 Å structure of a calmodulin-peptide complex. *Science.* 257:1251–1255.

81. van den Ent, F., J. Moller-Jensen, L. A. Amos, K. Gerdes, and J. Lowe. 2002. F-actin-like filaments formed by plasmid segregation protein ParM. *EMBO J.* 21:6935–6943.