# Study of Multidomain Protein Refinement by Different types of Scattering tools

**Indu Rathi**

Research Scholar, Singhania University, Jhunjhunu, Rajasthan, India

**ABSTRACT**: Determination of the 3D structures of multidomain proteins by solution NMR methods presents a number of unique challenges related to their larger molecular size and the usual scarcity of constraints at the interdomain interface, often resulting in a decrease in structural accuracy. In this respect, experimental information from small-angle scattering of X-ray radiation in solution (SAXS) presents a suitable complement to the NMR data, as it provides an independent constraint on the overall molecular shape. A computational procedure is described that allows incorporation of such SAXS data into the mainstream high-resolution macromolecular structure refinement. The method is illustrated for a two-domain 177-amino-acid protein, yS crystallin, using an experimental SAXS data set fitted at resolutions from ~200 A to ~30 A. Inclusion of these data during structure refinement decreases the backbone coordinate root-mean-square difference between the derived model and the high-resolution crystal structure of a 54% homologous yB crystallin from 1.96 ± 0.07 A to 1.31 ± 0.04 A. Combining SAXS data with NMR restraints can be accomplished at a moderate computational expense and is expected to become useful for multidomain proteins, multimeric assemblies, and tight macromolecular complexes..

---------------------------------------♦------------------------------------

## 1. INTRODUCTION

Determination of the three-dimensional structures of large proteins by solution NMR techniques presents a number of unique challenges. Increased line width resulting from slower rotational diffusion leads to a decrease in signal-to-noise ratio, increased resonance overlap, and larger uncertainty of the resonance positions. These effects decrease the number of observable NMR signals and complicate the process of their assignment. One way to address this problem is by combining $^{13}$C and $^{15}$N enrichment with perdeuteration, where the majority of $^{1}$H nuclei are replaced by the effectively NMR-invisible $^{2}$H.[1,2] When complemented by transverse relaxation-optimized spec- troscopy (TROSY)-based pulse sequence techniques, such labeling leads to a dramatic simplification of the NMR spectra, narrower resonance signals, and increased signal-to-noise ratios.[3] Perdeuteration, however, also has a downside: since it effectively makes sparse the set of NMR observables, it decreases the intrinsic information content of the NMR data. Additional difficulties arise due to the nonglobular nature of many multidomain proteins. Even though the conformations and relative orientations of the individual domains can be determined accurately by using backbone—backbone nuclear Overhauser effects (NOEs) and extensive sets of residual dipolar couplings (RDCs), relative positioning of the individual domains can remain challenging as protein perdeuteration eliminates the majority of the resonances necessary for defining the requisite side-chain-mediated interdomain NOE contacts.

Any source of experimental data that can compensate for the decrease in NOE restraint information associated with the application of NMR to large, multidomain proteins is therefore expected to be invaluable. In particular, information is needed that complements restraints derived from the common types of NMR data, including short-range interproton distances derived from NOEs,[4—7] dihedral angles derived from *J* couplings,[8,9] and orientations derived from residual dipolar couplings.[10],[11] It is well recognized that such complementary information is contained in the profiles of small-angle scattering of X-ray radiation by macromolecules in solution (SAXS).[12] Previously, SAXS data have been used in ad hoc calculations to complement NMR data in solving the solution structures of modular proteins (e.g., the Gla-EGF domain of the blood coagulation factor Xa protein[13] and a calmodulin/trifluoperazine complex[14]) essentially by evaluating which NMR-derived relative domain positions are in best agreement with the SAXS data or by a grid search for a 3D translation vector between the rigidly held domains. However, the potential for combining the two types of data has never been fully exploited directly in NMR structure calculation.

The SAXS intensity curve, recorded as a function of the scattering angle, is essentially a Fourier transform of the distribution of the interatomic distances within the macromol- ecule. Since the latter is known to encode both the overall molecular shape and the nonuniform distribution of the protein's atomic density,[15] incorporation of this information into macro- molecular structure refinement can compensate for the deficiency of the translational information derived from interdomain NOEs. Other advantages of using SAXS in the context of NMR- based structure determination are its independence of isotopic labeling, the high speed of data acquisition at the conditions that can be matched to those used for the solution NMR experiments, and smaller sample volumes ($\sim 15\ \mu\mathrm{L}$ per sample) compared to those required for NMR measurements. The main experimental challenges in applying SAXS methodology are the following: (i) sample conditions have to be carefully optimized to prevent aggregation, (ii) subtraction of the solvent contribution to the scattering must be done with high precision, and (iii) the sample can suffer radiation damage.

Here we demonstrate that direct incorporation of SAXS data in NMR structure calculation is readily feasible, and at moderate computational expense. The combination of NMR data, recently used for determining the solution structure of the eye lens protein $\gamma\mathrm{S}$ crystallin, with SAXS data results in considerably closer agreement with the X-ray structures of homologous members of the $\gamma$-crystallin family than the original NMR structure.

**Materials and Methods :**

**Protein Sample Preparation.** A uniformly $^{15}$N-enriched sample of $\gamma\mathrm{S}$ crystallin was used for collecting the SAXS data. Enrichment of the protein in $^{15}$N was used only because the sample initially was intended for NMR studies, and does not affect the protein stability or its scattering profile. Protein preparation details have been described elsewhere.[16] To minimize oxidation-induced dimerization through the Cys residues on the surface of the protein, the sample was dialyzed against 100 mL of buffer containing fresh reducing agent (dithiotreitol, DTT) for 6 h under the flow of N2 on-site, immediately prior to data acquisition. The sample composition was 9 mg/mL protein, 0.04% $NaN_3$, 5 mM DTT, 25 mM imidazole, pH 6.0. An aliquot of the dialysate was used to measure the solvent blank, which must be subtracted from the sample measurement in order to determine the scattering from the protein molecules alone. This same dialysate was also used for diluting the sample, to evaluate the concentration dependence of the SAXS profile.

**SAXS Data Acquisition and Processing.** Each $12\ \mu\mathrm{L}$ sample was centrifuged at $\sim 1000$ rpm into a glass capillary mounted on a brass holder, which was used to position the capillary precisely and reproducibly in the focused X-ray beam. Scattering data were acquired with the sample cooled to 291.4 K using the X-ray instrument at the University of Utah, described in a previous publication.[17] The instrument uses a sealed tube source (Cu Ka-edge giving 1.542 A wavelength) and a slit geometry with a one-dimensional position-sensitive detector. The sample-to-detector distance was 0.64 m, corresponding to an accessible *q* range of 0.0054—0.3192 $A^{—1}$. Individual detector channels were mapped onto the momentum transfer axis using the $50.1 \pm 0.1\text{Å}$ *d* spacing of the (100) reflection of the polycrystalline cholesterol myristate sample. To prevent oxidation of the sample by air during the measurement, N2 was flowing around the capillary throughout the experiment. Scattering data were acquired for 12 h per sample at two protein concentrations: 9.0 and 4.5 mg/mL. Data normalization, correction for the detector sensitivity, and subtraction of the solvent scattering were done as described previously.[17] Preliminary data analysis was done using Guinier formalism and P(r) analysis based on an indirect Fourier transform; it uses a sin(x)/x series expansion and is implemented in the program P_of_R that includes beam geometry corrections.[18] The P(r) analysis was also carried out using the program GNOM[19,20] which, along with the beam geometry corrections, utilizes a regularized indirect transform and thus avoids the potential for systematic oscillations in the calculated P(r). For the acquired $\gamma S$crystallin data, both programs gave essentially the same result, indicating that the scattering data are of good quality in that they have a robust P(r) solution, independent of the details of the Fourier transform. The contribution to the scattering arising from the hydration layer at the surface of the protein was calculated for a given structure by fitting the desmeared scattering data to the structure in question using the program CRYSOL.[21] The globbic correction was calculated from the structural coordinates using scattering profile simulation software written in-house, and available upon request from the authors.

**Structure Calculation Protocol.** $\gamma S$crystallin structure models were generated by a restrained molecular dynamics simulated annealing protocol using the CNS package.[22] The force field included the usual empirical energy terms: bonds, angles, improper angles, and a repulsive-only quartic nonbonded term with all van der Waals radii scaled down by a factor of 0.8, as well as a backbone—backbone hydrogen-bonding potential of mean force.[23] Additional terms included those for the NOEs, experimental dihedral angles, and RDCs, and were identical to those used previously for calculating the$\gamma S$crystallin structure in the absence of SAXS data (Protein Data Bank (PDB) entries 1ZWM and 1ZWO). The temperature was linearly decreased from 2000 K to 1 K in 200 stages of 200 steps each, with the $H^N$—N RDC force constant ramped up from 0.01 to 0.40 $kcal/Hz^2$. NOE and backbone dihedral angle force constants were fixed throughout the calculations at 50 $kcal/A^2$ and 10 $kcal/rad^2$, respectively. All statistics were extracted from the ensembles of 20 calculated structures, starting from the structures previously calculated and deposited in the absence of SAXS data. In all cases, reference calculations were run in exactly the same way, but with the SAXS data fit term inactivated. The original NMR structure of$\gamma S$crystallin was based primarily on backbone one-bond dipolar couplings, supplemented by a moderate number of easily accessible $H^N$—$H^N$ and CH3—CH3 NOE data. A total of 179 $H^N$—$H^N$ NOEs and 70 $CH_3$—$CH_3$ NOEs were available, 15 of them between the N- and C-terminal domains. The dipolar restraints include an extensive set of couplings recorded in two media, and comprise 291 N—H$^N$, 303 C—C$^a$, 273 N—C', and 246 C$^a$—C RDCs. Backbone dihedral angles $(\phi,\psi)$ are restrained by values derived from the previously described molecular fragment replacement (MFR) database search procedure,[16,24] which is based on the observed dipolarcouplings and yields a total of 318 torsion restraints. Restraints for $71\ \chi^1$ and 11 $\chi^2$ side-chain angles, extracted from $^3J_{C\gamma C'}$ and $^3J_{C\gamma N}$ couplings, were also used.

**Results and Discussion :**

**SAXS Data Analysis in the Context of High-Resolution Structure Refinement.** X-rays are scattered by electrons, and the intensity of the radiation scattered by the macromolecules in solution depends on the electron scattering density difference, or "contrast", between the macromolecule and the bulk solvent. An additional contribution to the scattering arises from a thin layer of solvent at the macromolecular surface which can have an electron density different from that of the bulk solvent. The existence of the latter hydration layer effect has been demonstrated in a number of experimental and computational stud- ies.21,25,26 In isotropic conditions, the scattering intensity is averaged over all orientations of the macromolecule with respect to the incident radiation beam. The scattering vector $q = 4\pi(\sin\theta)/\lambda$ denotes the momentum transfer between the incident beam of wavelength $\lambda$ and the radiation scattered at the angle $2\theta$. In the absence of macromolecular aggregation, the intensity of the scattered beam can be represented as[21]

$$I(q) = \langle |A_m(\mathbf{q}) - \rho_s A_s(\mathbf{q}) + \delta\rho A_l(\mathbf{q})|^2 \rangle_\Omega \qquad (1)$$

Here $\langle\rangle_\Omega$ denotes the solid angle average over all orientations of the momentum transfer vector q for the fixed norm q, $A_m$- (q), $A_s$(q), and $A_l$(q) are the scattering amplitudes of the macromolecule, solvent displaced by the macromolecular volume, and its hydration layer, respectively, and $\rho_s$ and $\delta\rho$ are the bulk solvent electron density (0.334 e/A$^3$) and the density of the hydration layer (0.00—0.07 e/A$^3$).[21] At a given orientation of the momentum transfer vector q with respect to the molecular frame, the scattering amplitude of the macromolecule is a Fourier transform of the atomic coordinates ry over its N atoms, weighted by the atomic X-ray form factors $f_j$:

$$A_m(\mathbf{q}) = \sum_{j=1}^{N} f_j(q)\exp(i\mathbf{q}\mathbf{r}_j) \qquad (2)$$

The scattering of the solvent displaced by the macromolecule can be approximated by placing dummy solvent atoms at all atomic positions within the macromolecule with the form factors given by[27]

$$g_j(q) = G(q)V_j \exp\left(-\frac{q^2 V_j^{2/3}}{4\pi}\right) \qquad (3)$$

Here, *Vy* are the volumes of the solvent displaced by each atom represented by the Gaussian spheres of previously tabulated[27] radii *ry*. The expansion factor G(q) is given by[21,25,26]

$$G(q) = \left(\frac{r_0}{r_m}\right)^3 \exp\left(-\frac{q^2({r_o}^2 - {r_m}^2)}{(36\pi)^{1/3}}\right) \qquad (4)$$

Here, *r*0 is the average atomic radius in the macromolecule and *r*m is the adjustable parameter that allows one to vary the average displaced solvent volume per atomic group. Here, we set $r_m$ = *r*0, which makes the expansion factor equal to one. The total scattering amplitude of the contrast between the macromolecule and the displaced solvent can then be conveniently expressed as the Fourier transform of the macromolecular coordinates weighted by the solvent-corrected form factors $f_j^s$:

$$\begin{aligned} A_{\mathrm{m}}(\mathbf{q}) - \rho_{\mathrm{s}} A_{\mathrm{s}}(\mathbf{q}) &= \sum_{j=1}^{N}[f_j(q) - \rho_{\mathrm{s}} g_j(q)] \exp(i\mathbf{q}\mathbf{r}_j) \\ &= \sum_{j=1}^{N} f_j^{\mathrm{s}}(q) \exp(i\mathbf{q}\mathbf{r}_j) \qquad (5) \end{aligned}$$

We will restrict our treatment to the range of $q < 1$ A$^{-1}$, where this approximate procedure can be expected to work reasonably well.

There are two common approaches to solid angle averaging over the exp('qr/) terms, one exploiting the favorable properties of their spherical harmonics expansion[21],[25],[26] and the other relying on application of the Debye formula.[28,29] Both involve a comparable computational overhead for proteins of up to $\sim 300$ residues. We chose the Debye formula for its mathematical simplicity, representing the spherical average in eq 1 as :

$$I(q) = \sum_{i=1}^{N} \sum_{j=1}^{N} f_i^{\mathrm{s}}(q) f_j^{\mathrm{s}}(q) \frac{\sin(qr_{ij})}{qr_{ij}} \qquad (6)$$

The quality of the fit between the experimental scattering data and those predicted from the model is described by the $\chi^2$ statistics over the set of $N_q$ experimental values

$$\chi^2 = \frac{1}{N_q - 1} \sum_{k=1}^{N_q} \left[ \frac{I_{\mathrm{expt}}(q_k) - c_k I_{\mathrm{calc}}(q_k)}{\sigma(q_k)} \right]^2 \qquad (7)$$

Here, $c_k$ are scattering vector-dependent correction factors described in more detail below and $a(q_k)$ are the uncertainties of each experimental data point $q_k$. Fitting SAXS data would thus involve simulation of the model-based scattering intensity $I_{calc}(q_k)$ for all $q_k$, correction of the latter by the $c_k$ factors, calculation of the $\chi^2$ statistics, and finally, differentiation of $\chi^2$ with respect to the current atomic coordinates to yield a set of atomic forces that aim to minimize $\chi^2$. When added to an empirical force field used in the molecular dynamics (MD)- based structure refinement, these forces should allow a refinement against SAXS data in combination with other data sources (in this case, a set of NMR-generated restraints). The gradient of the $\chi^2$ with respect to the atomic coordinates $r_j$ can be expressed as

$$\nabla_{r_j}[\chi^2] \approx \sum_{k=1}^{N_q} \frac{I_{\mathrm{expt}}(q_k) - c_k I_{\mathrm{calc}}(q_k)}{\sigma_k^{\,2}} \sum_{i=1}^{N} \sum_{j \neq i}^{N} f_i^{\mathrm{s}}(q_k) f_j^{\mathrm{s}}(q_k) \times \left[ \cos(q_k r_{ij}) - \frac{\sin(q_k r_{ij})}{q_k r_{ij}} \right] \frac{\mathbf{r_{ij}}}{r_{ij}^{\,2}} \qquad (8)$$

Hence, fitting SAXS data involves evaluation of eqs 6—8 at each step of molecular dynamics/energy minimization. Because the number of operations necessary for these calculations scales as $NqN^2$, it is clear that one problem that has been preventing incorporation of SAXS data into structure refinement is its enormous computational overhead. For example, calculation of the $\chi^2$ and its gradients takes tens of seconds of CPU time on a modern Pentium-class processor per step, for proteins between 100 and 200 residues in length. Since MD trajectories commonly used in high-resolution structure refinement may involve $10^4$— $10^5$ such steps, the challenges are quite apparent.

The solution to this problem is hinted at by the form of the $N_qN^2$ expression: a suitable approximation to eqs 6—8 with smaller values of $Nq$ and $N$ will alleviate the computational burden. Starting with $N^2$-dependent terms, it is known that the shapes of the spherically averaged scattering form factors of small, closely proximal sets of atoms do not show a pronounced dependence on the exact atomic geometries below $\sim 3$ Å resolution.[30] The resulting "globbic approximation", in which an all-atom representation of the macromolecular structure is coarse-grained into a smaller number of spatially proximal "globs", has been widely used in the interpretation of the low- resolution X-ray crystallographic[31] and SAXS[1928] data. Following this strategy, we have split protein structures into sets of small fragments, each involving 3—9 heavy atoms, along with their associated H's (see the Supporting Information for the definition of the "globs"). We have then recalculated the spherically averaged scattering form factors for each glob as

$$^{\mathrm{glob}}f_k^{\mathrm{s}}(q) = \left[\sum_{i=1}^{N}\sum_{j=1}^{N} f_i^{\mathrm{s}}(q) f_j^{\mathrm{s}}(q) \frac{\sin(qr_{ij})}{qr_{ij}}\right]^{1/2} \qquad (9)$$

One can then approximate the scattering intensity curve with the sum in eq 6 running over the set of globs, positioned at the coordinates weighted by the atomic electron number counts within each glob, and using the globbic form factors instead of the atomic ones. Since our specification reduces *N* input heavy atoms into approximately N/3 globs, the required CPU time is reduced by about an order of magnitude. The procedure, however, has a drawback: the approximated scattering intensity curves show small but systematic differences with respect the "exact" ones, obtained from all-atom calculations. We address this problem via an approach used by others:[31] derivation of "globbic" correction factors *ck = c(qk)* as ratios between the "exact" scattering curves and the globbically approximated ones. Figure 1 shows the average and standard deviation of this correction, calculated over a large set of protein structures in the 100—200 residue size range. Application of such a correction will decrease the systematic errors of our approximation to values comparable to the error bars indicated within the figure.

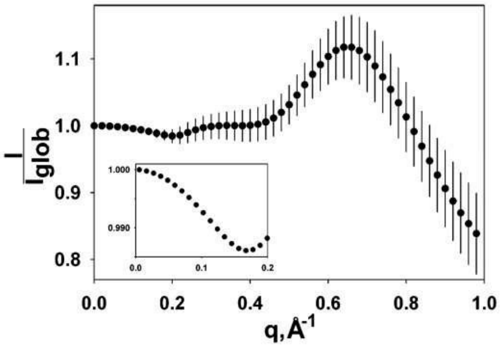

***Figure 1.*** Globbic correction factor calculated as the ratio between the atomic and globbic scattering curves, *I(q)* and $I_{\mathrm{glob}}(q)$, respectively. The mean and standard deviation of the curve points are calculated on the basis of 538 single-chain protein X-ray structures of 100—200 residues length, solved at resolutions of 1.8 A or better. The calculations were carried out according to eqs 6 and 9. The inset shows the average correction factor calculated from the $\gamma S$ crystallin models used in the final round of structure refinement.

Notice that since our globs are smaller than the "dummy residues" usually employed in SAXS data analysis, the average correction factors and their variances are smaller than the ones obtained in those approaches (compare to Figure 2 of ref 19). In fact, we have adjusted the size and composition of the globs to provide a conservative compromise between the computational speed-up and the magnitudes of the systematic errors resulting from the approximate nature of the calculation. The shape and overall features of the globbic correction curve are largely independent of the size and secondary structure content of the protein, while showing a pronounced dependence on the glob size, especially in the higher resolution range (see Supporting Information for details).

In practice, these correction factors are calculated from the current structural model, and re-estimated after each successive cycle of structure refinement until convergence is reached. Such a procedure will, in general, ensure that the approximated globbic correction curve approaches the exact one as the refined structure approaches the correct model. The calculated scattering intensity curves are also corrected for the effect of the bound solvent layer using CRYSOL,[21] taking as input the entire family of structures prior to every cycle of structure refinement and fitting the bound solvent density as the only adjustable parameter.

The second part of our strategy involves reducing Nq, the number of experimental points to be fitted. For proteins of up to a few hundred residues, the maximum curvature of the simulated scattering curves, ca. $10^{-2}$ A $^{-}$ i s much smaller than the scattering vector step of the oversampled experimental data (typically ca. $10^{-3}$ $A^{-11}$). Reduction of the fitted data set to fewer points within the same *q* interval is thus expected to speed-up the calculation by an amount proportional to the ratio of the number of points in the original data to that in the "sparsened" data set. If the separation in *q* between the sparsened data points is substantially smaller than the distance between the features of the scattering curve, sparsening is not expected to have any detrimental effects on the accuracy of the data representation. We have performed a regularized fit of the oversampled, desmeared experimental data set using the package GNOM[20 32] and sparsened the smoothened data fit by a factor of 8. The combination of these two procedures results in an overall speedup factor of $\sim 80,$placing the time for a single-point SAXS pseudo-energy/forces calculation to less than $\sim 1/3$ of a second for a protein of up to $\sim 180$residues, when fitting up to 30 SAXS data points on a 2.8 GHz Pentium 4 processor. This gain makes it possible to conduct regular-length MD structure refinement in a reasonable amount of time (ca. 6 h per structure for 40 000 MD steps). The SAXS data fitting module was coded into the CNS structure refinement package[22] with the corresponding energy term introduced by the "SAXS" keyword.

**Application to $\gamma\mathrm{S}$ Crystallin.** We demonstrate the utility of the solution scattering data in NMR structure refinement of murine $\gamma\mathrm{S}$crystallin, a two-domain eye lens protein of 177 residues. The N- and C-terminal domains are topologically similar, each consisting of two four-strand $\beta$-sheets arranged in Greek key motifs, linked by a Tyr corner. The entire protein shares 54% sequence identity with bovine $\gamma\mathrm{B}$crystallin, for which a $1.1$ Å resolution X-ray structure is available (PDB code 1AMM[33]), and 50% sequence identity with human $\gamma\mathrm{D}$crystallin (PDB code 1HK0[34]). In addition, a crystal structure is available for a dimer formed by the C-terminal domains of bovine $\gamma\mathrm{S}$ crystallin (PDB code 1A7H[35]). The primary sequence of $\gamma\mathrm{S}$ crystallin can be aligned to these entries without any gaps or insertions within each individual domain.

The NMR structure for $\gamma\mathrm{S}$crystallin was recently determined by molecular fragment replacement (MFR) methodology,[24] using primarily dipolar couplings as input restraints, supplemented by small numbers of $H^N$—$H^N$ and CH3—CH3 NOE restraints.[16] The two globular domains of the recent NMR structure of $\gamma\mathrm{S}$ crystallin are very similar to those seen in the homologous $\gamma\mathrm{B}$ crystallin (backbone root-mean-square deviation (rmsd) 0.63 and 1.09 Å for the N- and C-terminal domains, respectively). The relative orientation of the two domains in $\gamma\mathrm{S}$crystallin is also very similar to that seen in other crystallin structures, but the two domains are farther apart in the NMR structure, presumably as a result of the scarcity of interdomain restraints. This situation is encountered more frequently, in particular in protein—protein complexes, and in larger proteins where interdomain NOEs tend to be relatively sparse, but relative orientations of domains are tightly defined by RDCs.[36] [37] Therefore, the SAXS data present an ideal complement for determining an accurate solution structure of such systems.

The SAXS data for $\gamma$S crystallin at 4.5 mg/mL protein concentration were minimally affected by aggregation, as determined by the linearity of the Guinier plot (see Supporting Information) and P(r) analysis. The latter yields a gyration radius ($R_g$) value of $18.3 \pm 0.2$ Å, a maximum linear dimension ($D_{max}$) of 54—57 Å, and an estimated molecular volume of $(25.2 \pm 0.7) \times 10^3$ Å$^3$, approximated from the total intensity under the measured scattering profile and using the Porod invariant.[38] The same parameters determined using the 1AMM, 1A7H, and 1HK0 crystal structures and the program CRYSOL[21] are $R_g$ = 16.6—16.8 Å, $D_{max}$ = 55.2—56.5 Å, and a molecular volume of $(25.5-25.8) \times 10^3$ Å$^3$.

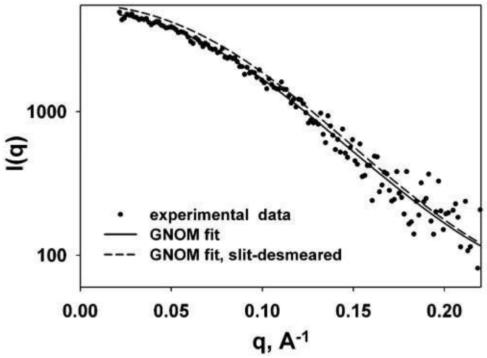

***Figure 2.*** Experimental scattering data recorded for the 4.5 mg/mL $\gamma$S crystallin sample. The solid line shows regularized data fit from the GNOM program. The dashed line corresponds to the slit-desmeared data fit. A total of 16 points of this curve, equally spaced between $\sim 0.02$ and $\sim 0.22$ Å$^{-1}$, are subsequently used for structure calculation.

The observed difference in *R*g is likely to be a consequence of a thin surface layer of solvent with a density higher than that of the bulk solvent, a phenomenon often leading to an increase of the apparent SAXS-extracted $R_g$ values by 1—2 Å with respect to the numbers calculated from the atomic coordinates. A weak tail is seen in the P(r) distribution that appears to have a $D_{\max}$ of ca. 80 Å, which likely reflects a small amount of dimerized protein in the sample volume. The presence of seven reduced Cys residues in $\gamma$S crystallin, of which surface-exposed $Cys^{24}$ and $Cys^{26}$ are particularly reactive, promotes dimerization and formation of higher-order multimers under oxidizing conditions. *I*(0) analysis of the data, using lysozyme as a standard, indicates that the dimers account for less than 8.5% of the total protein. The raw data as well as the regularized GNOM fits are shown in Figure 2. Even though the recorded scattering intensity extends up to 0.32 Å$^{-1}$, the uncertainty in our data precludes interpretation beyond about 0.22 Å$^{-1}$. The increased uncertainty is due in part to the fact that the SAXS instrument used has a one-dimensional detector and hence captures an increasingly smaller percentage of the solid angle of the circularly averaged scattering pattern at larger angles; a much higher signal-to-noise ratio can be attained using a synchrotron source coupled with an area detector, providing the sample can withstand the high radiation levels.

A total of five cycles of structure refinement were necessary to make globbic and surface solvent layer corrections consistent with the ensemble of refined structures. The density of the bound solvent layer, assumed to be 3.5 Å thick, was determined from CRYSOL fits to be 0.025 e/Å$^3$ higher than the bulk solvent density, which is within the expected range for a typical protein in solution.

The accuracy of the atomic coordinates of the refined models was evaluated with respect to the high-resolution X-ray structures of$\gamma\mathrm{B}$, $\gamma\mathrm{D}$,and C-terminal $\gamma\mathrm{S}$ crystallins (PDB entries 1AMM, 1A7H, and 1H0K). The $\gamma\mathrm{B}$and$\gamma\mathrm{D}$crystallins share ca. 50% sequence identity with $\gamma\mathrm{S}$crystallin, and 74%with one another. With a two-domain backbone rmsd of 0.69 Å,the crystal structures of$\gamma\mathrm{B}$and $\gamma\mathrm{D}$ crystallins exhibit very close similarity, despite crystallization in two different space groups. When comparing relative domain positions in $\gamma\mathrm{B}$and $\gamma\mathrm{D}$ (keeping their N-terminal domains superimposed), the orientations of their C-terminal domains differ primarily by a 5.5° rotation and exhibit no detectable translation. The packing at the hydrophobic interface in the homodimer of the C-terminal $\gamma\mathrm{S}$crystallin domain is similarly tight, but shows a 23° rotation relative to $\gamma\mathrm{B}$. In contrast, in our previously determined solution structure of$\gamma\mathrm{S}$crystallin, the backbone rmsd relative to $\gamma\mathrm{B}$and $\gamma\mathrm{D}$ is dominated by translation, not by relative domain orientation, and presumably results from insufficient interdomain NOE restraints.[16] Therefore, this backbone rmsd presents a reasonable measure for the error in the relative position of the two domains of$\gamma\mathrm{S}$crystallin.

The dependence of small-angle scattering intensity on the square of the molecular weight of the scattering particle results in a scattering profile that is quite sensitive to small amounts of aggregation. In contrast, NMR is relatively insensitive to minor degrees of aggregation in the sample. Thus, combining NMR and scattering data could be problematic if the procedure were intolerant to even weak degrees of self-association. Considering that$\gamma\mathrm{S}$crystallin has a tendency to self-associate, as judged by the steeper than expected increase in rotational correlation time with volume fraction, and to form covalent homodimers through oxidation of the solvent-exposed Cys$^{24}$ and Cys$^{26}$ residues, it presents a challenging case for SAXS refinement. Therefore, the fact that we obtained a considerable improvement in structural accuracy for this rather challenging system bodes well for the future utility of this technique. It is also encouraging that significant gains in structural accuracy can be made even with the relatively modest statistical quality of our SAXS data, which were obtained using a simple laboratory-based instrument that uses a sealed tube X-ray source. Scattering profiles extending to much higher angles and at much higher signal-to-noise ratios can be recorded at synchrotrons for favorable systems, such as larger proteins and nucleic acids.[40]

Our structure refinement procedure is based on the assumption of a single, well-defined conformation. However, it is important to bear in mind that SAXS data represent an average over all conformations sampled by the molecule in solution. In the application to $\gamma\mathrm{S}$crystallin, the assumption of a single well- defined conformation is supported by a variety of NMR data, including $^{15}$N backbone dynamics measurements and the indistinguishable values of the alignment tensors of the two domains. However, there is no a priori reason that prevents application of the SAXS refinement procedure to a multicon- former refinement of a more dynamic complex.

Another issue of potential interest is whether including SAXS data in the refinement, as done in the current study, has any advantages over calculating a family of structures and then selecting from these the subset with the lowest $\chi^2$ of the SAXS data fit, a task that can easily be performed with existing software.[13,14] We have generated a family of 166 structures without inclusion of SAXS data, and evaluated SAXS $\chi^2$ on those models (see Supporting Information). Our results indicate that, while selection by the lowest SAXS$\chi^2$ will lower the rmsd to 1AMM, the decrease is considerably smaller than when the SAXS data are fitted directly. This outcome results in part from the commonly used "repulsive-only" nonbonded interactions, and underscores the limitations in providing sufficient sampling of conformational space during the structural refinement, which can be overcome by including the SAXS data as restraints in the structure calculation.

**Concluding Remarks :**

In this study we have demonstrated the utility of solution X-ray scattering data as a component of high-resolution NMR structure refinement. The obtained improvements in accuracy are very encouraging, particularly given the limited effective resolution range of only up to $\sim 30$ Å spanned by our acquired scattering data. SAXS data present an ideal complement to NMR data sets rich in orientational restraints, such as those contained in residual dipolar couplings, but lacking a large number of accurate translational restraints, such as NOEs. Use of the SAXS data clearly will be most advantageous for defining the solution structure of larger macromolecules, where the number of restraints per residue tends to be sparse, but where dipolar couplings are still readily accessible. Higher informational content within the same resolution range and higher signal-to- noise ratios for SAXS data when applied to these systems is well suited to offset the decrease of the density of the NMR- based structural constraints.[41—43]

To date, the usage of SAXS data in structural biology has mainly been limited to (i) de novo low-resolution shape reconstruction, (ii) testing previously derived high-resolution structural models, and (iii) rigid-body refinement of multiunit macromolecular assemblies. With the substantial improvements in the formalism connecting the observed data to the underlying structural model that has occurred in the past few years, this situation is likely to change. The direct fitting approach described in the current study is intended to facilitate a more routine usage of this key data source during macromolecular structure refinement.

**References**

(1) Grzesiek, S.; Anglister, J.; Ren, H.; Bax, A. *J. Am. Chem. Soc.* **1993**, *115*, 4369-4370.

(2) Tugarinov, V.; Hwang, P. M.; Kay, L. E. *Annu. ReV. Biochem.* **2004**, *73*, 107-146.

(3) Salzmann, M.; Wider, G.; Pervushin, K.; Senn, H.; Wuthrich, K. *J. Am. Chem. Soc.* **1999**, *121*, 844-848.

(4) Wu¨thrich, K. *NMR of Proteins and Nucleic Acids*; John Wiley & Sons: New York, 1986.

(5) Kaptein, R.; Boelens, R.; Scheek, R. M.; van Gunsteren, W. F. *Biochemistry* **1988**, *27*, 5389-5395.

(6) Clore, G. M.; Gronenborn, A. M. *Crit. ReV. Biochem. Mol. Biol.* **1989**, *24*, 479-564.

(7) Wagner, G. *J. Biomol. NMR* **1993**, *3*, 375-385.

(8) Biamonti, C.; Rios, C. B.; Lyons, B. A.; Montelione, G. T. *AdV. Biophys. Chem.* **1994**, *4*, 51-120.

(9) Bax, A. *Methods Enzymol.* **1994**, *239*, 79-125.

(10) Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9279-9283.

(11) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111-1114.

(12) Svergun, D. I.; Koch, M. H. J. *Rep. Prog. Phys.* **2003**, *66*, 1735-1782.

(13) Sunnerhagen, M.; Olah, G. A.; Stenflo, J.; Forsen, S.; Drakenberg, T.; Trewhella, J. *Biochemistry* **1996**, *35*, 11547-11559.

(14) Mattinen, M. L.; Paakkonen, K.; Ikonen, T.; Craven, J.; Drakenberg, T.; Serimaa, R.; Waltho, J.; Annila, A. *Biophys. J.* **2002**, *83*, 1177-1183.

(15) Koch, M. H. J.; Vachette, P.; Svergun, D. I. *Q. ReV. Biophys.* **2003**, *36*, 147-227.

(16) Wu, Z.; Delaglio, F.; Wyatt, K.; Wistow, G.; Bax, A. *Protein Sci.* **2005**, *14*, 3101-3114.

(17) Heidorn, D. B.; Trewhella, J. *Biochemistry* **1988**, *27*, 909-915.

(18) Moore, P. B. *J. Appl. Crystallogr.* **1980**, *13*, 168-175.

(19) Svergun, D. I.; Petoukhov, M. V.; Koch, M. H. J. *Biophys. J.* **2001**, *80*, 2946-2953.

(20) Svergun, D. I. *J. Appl. Crystallogr.* **1992**, *25*, 495-503.

(21) Svergun, D.; Barberato, C.; Koch, M. H. J. *J. Appl. Crystallogr.* **1995**, *28*, 768-773.

(22) Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. *Acta Crystallogr. D, Biol. Crystallogr.* **1998**, *54*, 905-921.

(23) Grishaev, A.; Bax, A. *J. Am. Chem. Soc.* **2004**, *126*, 7281-7292.

(24) Kontaxis, G.; Delaglio, F.; Bax, A. *Methods Enzymol.* **2005**, *394*, 42-78.

(25) Merzel, F.; Smith, J. C. *Acta Crystallogr. D, Biol. Crystallogr.* **2002**, *58*, 242-249.

(26) Svergun, D. I. *Biophys. J.* **1999**, *76*, 2879-2886.

(27) Fraser, R. D. B.; Macrae, T. P.; Suzuki, E. *J. Appl. Crystallogr.* **1978**, *11*, 693-694.

(28) Chacon, P.; Moran, F.; Diaz, J. F.; Pantos, E.; Andreu, J. M. *Biophys. J.* **1998**, *74*, 2760-2775.

(29) Walther, D.; Cohen, F. E.; Doniach, S. *J. Appl. Crystallogr.* **2000**, *33*, 350-363.

(30) Guo, D. Y.; Blessing, R. H.; Langs, D. A.; Smith, G. D. *Acta Crystallogr. D, Biol. Crystallogr.* **1999**, 55, 230-237.

(31) Guo, D. Y.; Blessing, R. H.; Langs, D. A. *Acta Crystallogr. D, Biol. Crystallogr.* **2000**, *56*, 1148-1155. (32) Svergun, D. I. *Biophys. J.* **1991**, *24*, 485-592.

(33) Kumaraswamy, V. S.; Lindley, P. F.; Slingsby, C.; Glover, I. D. *Acta Crystallogr. D, Biol. Crystallogr.* **1996**, *52*, 611-622.

(34) Basak, A. K.; Bateman, O.; Slingsby, C.; Pande, A.; Asherie, N.; Ogun, O.; Benedek, G. B.; Pande, J. *J. Mol. Biol.* **2003**, *328*, 1137-1147.

(35) Basak, A. K.; Kroone, R. C.; Lubsen, N. H.; Naylor, C. E.; Jaenicke, R.; Slingsby, C. *Protein Eng.* **1998**, *11*, 337-344.

(36) Clore, G. M. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 9021-9025.

(37) Tugarinov, V.; Choy, W. Y.; Orekhov, V. Y.; Kay, L. E. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 622-627.

(38) Glatter, O.; Kratky, O. *Small-Angle X-ray Scattering*; Academic Press: New York, 1982.

(39) Sokolova, A. V.; Volkov, V. V.; Svergun, D. I. *J. Appl. Crystallogr.* **2003**, *36*, 865-868.

(40) Zuo, X. B.; Tiede, D. M. *J. Am. Chem. Soc.* **2005**, *127*, 16-17.

(41) Tugarinov, V.; Kay, L. E. *J. Mol. Biol.* **2003**, *327*, 1121-1133.

(42) Lukin, J. A.; Kontaxis, G.; Simplaceanu, V.; Yuan, Y.; Bax, A.; Ho, C. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 517-520.

(43) Williams, D. C.; Cai, M. L.; Clore, G. M. *J. Biol. Chem.* **2004**, *279*, 1449-1457.

(44) Koradi, R.; Billeter, M.; Wuthrich, K. *J. Mol. Graph.* **1996**, *14*, 51-55.