# Improved fitting of solution X-ray scattering data to macromolecular structures and structural ensembles by explicit water modeling

Alexander Grishaev,[a,*] Liang Guo,[b] Thomas Irving,[b] and Ad Bax[a,*]

[a] *Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892,*

[b] *Biophysics Collaborative Access Team, CSRRI ,BCPS Dept., Illinois Institute of Technology, Chicago, IL 60616*

# SUPPORTING INFORMATION

**Full reference 18:** Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J. M.; Liu, G. H.; Eletsky, A.; Wu, Y. B.; Singarapu, K. K.; Lemak, A.; Ignatchenko, A.; Arrowsmith, C. H.; Szyperski, T.; Montelione, G. T.; Baker, D.; Bax, A. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 4685-4690.

## Details of the AXES fitting procedure

**Water box generation.** The coordinates of the molecular solvent surrounding the macromolecules were taken from 200 snapshots of a constant (N,V,T) molecular dynamics run on a 74.524 Å cube of TIP5P water[1] (a total of 13824 molecules) at 25 ºC. Electrostatic interactions were modeled with a particle-mesh Ewald formalism[2] and subject to periodic boundary conditions. Switching and cut-off distances of 13 and 15 Å were used for non-bonded interactions. Accumulation of such snapshots was done after 100 ps system equilibration followed by production runs with each stored conformation separated by 10 ns from the previous one. The NAMD package[3] was used for the molecular dynamics simulation.

**Selection of the displaced and surface solvent molecules.** For a set of 50 different water boxes, the macromolecule is placed in its center. Displaced and surface solvent water molecules were selected for each individual water box by selecting those waters whose O atoms either clash with the macromolecule or form a non-clashing set of less than $r_{vdW} + r_{wat} + 3$ Å away from it. Here, $r_{vdW}$ is a set of Bondi van der Waals radii[4] and $r_{wat} = 1.4$ Å is the effective radius of the water molecule. In cases where the maximum macromolecular dimension exceeds 74.524 Å, additional water box images are positioned to obtain a complete coverage of the structure. Water molecules that are closer than $r_{vdW}$ from at least one atom of the macromolecules are accumulated in the displaced set and

those further than $r_{vdW} + r_{wat}$ into the surface solvent set. The waters situated between $r_{vdW}$ and $r_{vdW} + r_{wat}$ from the macromolecule are placed into either the displaced or surface water set by a stochastic procedure that aims to mimic the continuous repulsive potential between the macromolecule and the solvent. For these waters, the probability of placement in the surface set is calculated over all macromolecular atoms that are closer than $r_{vdW} + r_{wat}$ as

$$P_{surf} = \prod_j \left( \frac{r_j - r^j_{vdW}}{r_{wat}} \right)^{\gamma}$$

where $r_j$ is the distance between the protein atom and the water O atom, and the product extends over all protein atoms, $j$, for which $r - r_{vdW} < r_{wat}$. The exponent $\gamma$ was empirically adjusted to 0.10, such as to best reproduce the proper specific volumes of the proteins. The individual contributions are bracketed by 0 for $r_j < r^j_{vdW}$ and 1 for $r_j > r^j_{vdW} + r_{wat}$. The candidate water molecules are then stochastically partitioned into the displaced and surface sets according to the calculated $P_{surf}$ and $P_{disp} = 1\text{-}P_{surf}$ values via the Metropolis algorithm.[5] As a post-processing procedure, surface water molecules with less than 2 neighbors within 3.8 Å in the surface set are transferred into the displaced set. This procedure, performed iteratively until convergence is reached, removes the effects of small cavities possible within the macromolecule with the employed set of the atomic radii.

**Calculation of the predicted scattering intensity functions.**

The scattering intensity predicted from the macromolecular coordinates is calculated as the linear combination of the 6 elementary scattering functions averaged over angular orientations, macromolecular conformers, and molecular solvent configurations for a given electron density contrast $\delta\rho$ of the surface solvent layer, assumed to have a thickness of 3 Å:

$$I_{pred}(q) = \left\langle \left\langle \left\langle \left| \mathbf{F}_{mol} - \mathbf{F}_{disp} + \delta\rho \mathbf{F}_{surf} \right|^2 \right\rangle_{\Omega} \right\rangle_{solv} \right\rangle_{ens} =$$

$$I_{mol\text{-}mol}(q) + I_{disp\text{-}disp}(q) + \delta\rho^2 I_{surf\text{-}surf}(q) -$$
$$2I_{mol\text{-}disp}(q) + 2\delta\rho\left[ I_{mol\text{-}surf}(q) - I_{disp\text{-}surf}(q) \right]$$

The elementary scattering functions were accumulated by averaging over 1589 equi-spaced directions for each q-vector, corresponding to the 16th order Fibonacci number grid, and over 50 independent solvent configurations. Solvent scattering was calculated over both O and H atoms using form factors corrected for the atomic charge transfer to match the $H_2O$ liquid-phase dipole moment as specified by Sorenson et al.[6]

## Sample preparation

The B3 domain of Igg-binding protein G was expressed and purified as described previously[7] and ubiquitin, lysozyme and cytochrome C were purchased from Sigma-Aldrich. Proteins were dissolved in buffers composed of 150 mM NaCl, 40 mM Na acetate, 0.05% $NaN_3$, 5 mM DTT at pH 5.5 for GB3 and ubiquitin; the same buffer

composition at pH 4.3 was used for lysozyme; and 150 mM NaCl, 10 mM TRIS, 0.05% $NaN_3$, 5 mM DTT at pH 7.0 was used for cytochrome C. The samples were extensively dialyzed (>24 h) against degassed buffer, using membranes with molecular weight cut-off values of 3500 kDa. Samples were filtered through a 0.22 μm membrane and diluted to the stock concentration of 10 mg/mL. Concentrations of the proteins were measured at 50-fold dilutions in a solution of 6 M Guanidinium-HCl by UV absorption using extinction coefficients ($\varepsilon_{278}$=1.85 mL $mg^{-1}$ for GB3, $\varepsilon_{280}$=0.149 mL $mg^{-1}$ for ubiquitin, $\varepsilon_{280}$=2.55 mL $mg^{-1}$ for lysozyme, calculated from the proteins' sequences and free amino acid values by the ExPASy server[8]; and $\varepsilon_{550}$=2.29 mL $mg^{-1}$ for cytochrome C under native conditions). The samples and the dialysis-matched buffers were sealed and stored under nitrogen at 4 ºC until data collection (1-3 days).

**Experimental data collection and processing.** Experimental solution scattering data were acquired during several data collection sessions, spanning the period from 2007 to 2009, at the BIOCAT and BESSRC facilities at the Advanced Photon Source, Argonne National Laboratory (beamlines 18-ID and 12-IDC, respectively). Aviex and Gold mosaic CCD area detectors were used for data collection, respectively. Scattering intensities were recorded at 25 ºC in two geometries, with sample-to-detector distances of 200-400 cm and 35-50 cm in order to cover both SAXS and WAXS q-ranges with the total merged data range extending from ~0.008 $Å^{-1}$ to ~2.3 $Å^{-1}$. Data collections were done sequentially with the buffer followed by the sample. With both setups, sample volumes of 100-150 μL were flowing during data collection in order to minimize radiation damage. Each data collection included 20-40 individual frames with exposures of 0.1-0.6 sec. Data collection times were optimized to prevent both radiation damage and detector saturation near the beam stop and the primary water ring at ~ 2.0 $Å^{-1}$. Frames that showed systematic deviations from the majority of the data were discarded and the remaining sets were averaged. Before averaging, the data from the individual frames were scaled by the recorded incident intensities and transmission coefficients. Scattering intensity profiles from the capillaries were subtracted from both sample and buffer profiles, which were then subtracted from each other with the buffer scaled by the solvent volume fraction $\alpha = 1\text{-}c_{mg/mL}*7.425\cdot10^{-4}$ :

$$I_{\exp t}(q) = (I_{sam}(q) - I_{cap}(q)) - \alpha(I_{buf}(q) - I_{cap}(q))$$

For all four proteins, concentration series were recorded in SAXS configurations at 2.5, 5.0 and 10.0 mg/mL. In all cases, 2.5 and 5.0 mg/mL proved to be indistinguishable by Guinier analysis and the higher concentration data were used for further analysis. All data sets were monitored for aggregation via P(r), $R_{gyr}$, and I(0) analysis;[9] none was detected.

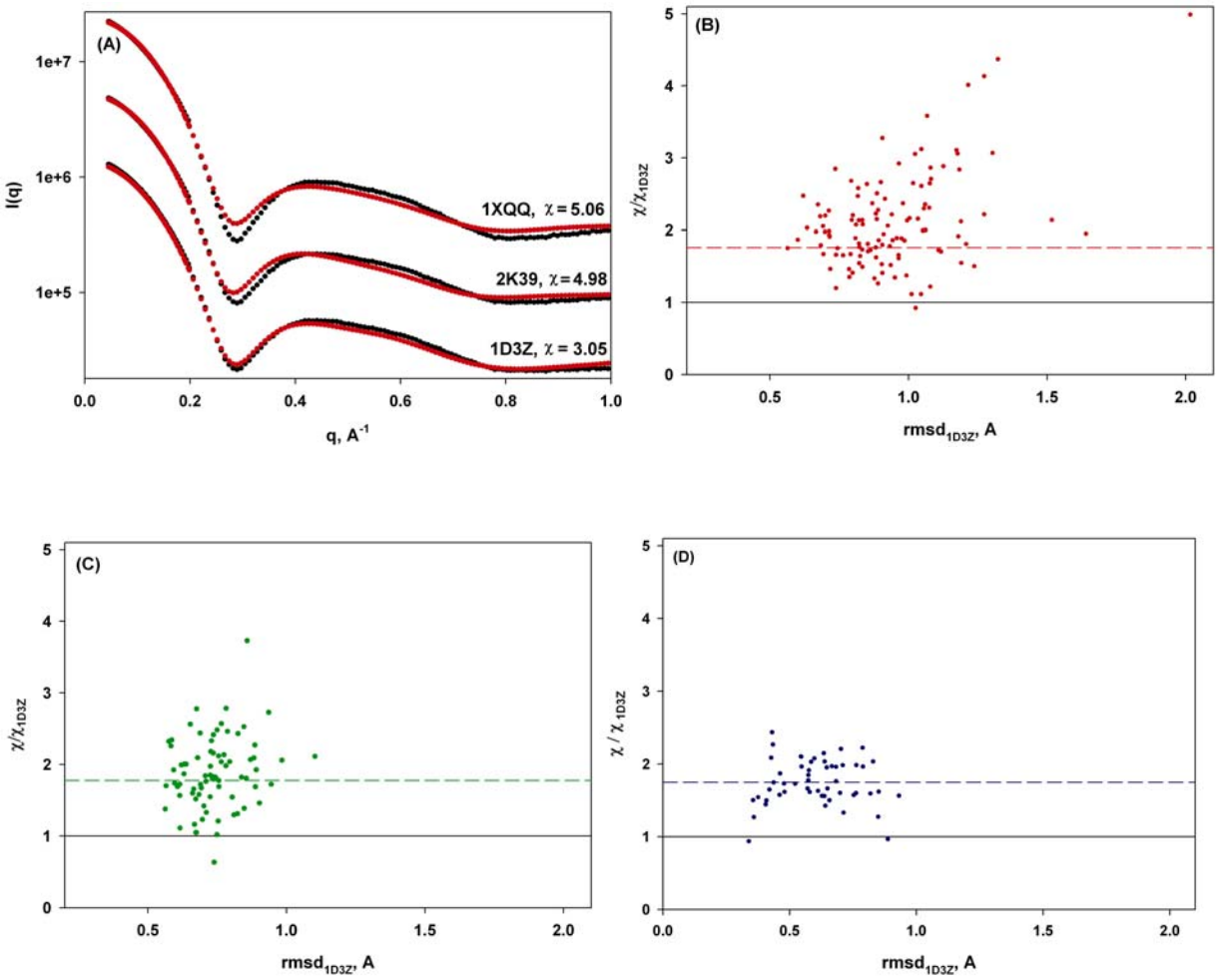


**Figure S1.** SAXS data fit for various models of ubiquitin. (A) AXES fits for 1XQQ,[10] 2K39,[11] and 1D3Z[12] ensembles of structures are offset for clarity and labeled with the PDB code of the fitted structures. Experimental data are shown in black and calculated data in red. For 1D3Z, the fit to the lowest energy structure (model 1) is shown. (B, C, D)**:** Quality of the fit of the individual structures belonging to the ensembles of Protein Data Bank entries 1XQQ (B, red), 2K39 (C, green), and the 53 full-length ubiquitin X-ray models deposited in the PDB (D, blue). Results in (B-D) are shown as functions of the backbone rmsd of each individual model (over residues 1-71) relative to the first model of the 1D3Z ensemble. Ratios of the best-fitted $\chi$ value of the individual structure over the $\chi$ value of the 1D3Z structure are plotted on the vertical scale. Normalized $\chi$ values obtained when fitting all ensemble members simultaneously are shown as dashed lines. The X-ray structure ensemble includes the following PDB depositions and chains: 1AAR:A, 1AAR:B, 1CMX:B, 1F9J:A, 1NBF:C, 1NBF:D, 1S1Q:D, 1TBE:A, 1UBI:A, 1UBQ:A, 2AYO:B, 2G45:B, 2G45:E, 2GMI:C, 2HD5:B, 2JF5:B, 2O6V:B, 2O6V:C, 2O6V:D, 2O6V:F, 2O6V:G, 2QHO:C, 2WDT:B, 2WWZ:A, 2ZNV:B, 2ZNV:C, 2ZNV:E, 3A1Q:B, 3A1Q:E, 3A9J:A, 3A9J:B, 3A9K:A, 3A9K:B, 3A33:B, 3CMM:B, 3CMM:E, 3DVG:Y, 3DVN:V, 3DVN:Y, 3H7P:B, 3HM3:A, 3HM3:B, 3HM3:C, 3HM3:D, 3JSV:A, 3JSV:B, 3JVZ:X, 3JVZ:Y, 3JW0:X, 3JW0:Y, 3M3J:A, 3M3J:C, 3M3J:E. AXES fits were performed averaging over 10 water boxes for each structure, a Fibonacci grid order of 16, and default settings for all other parameters.

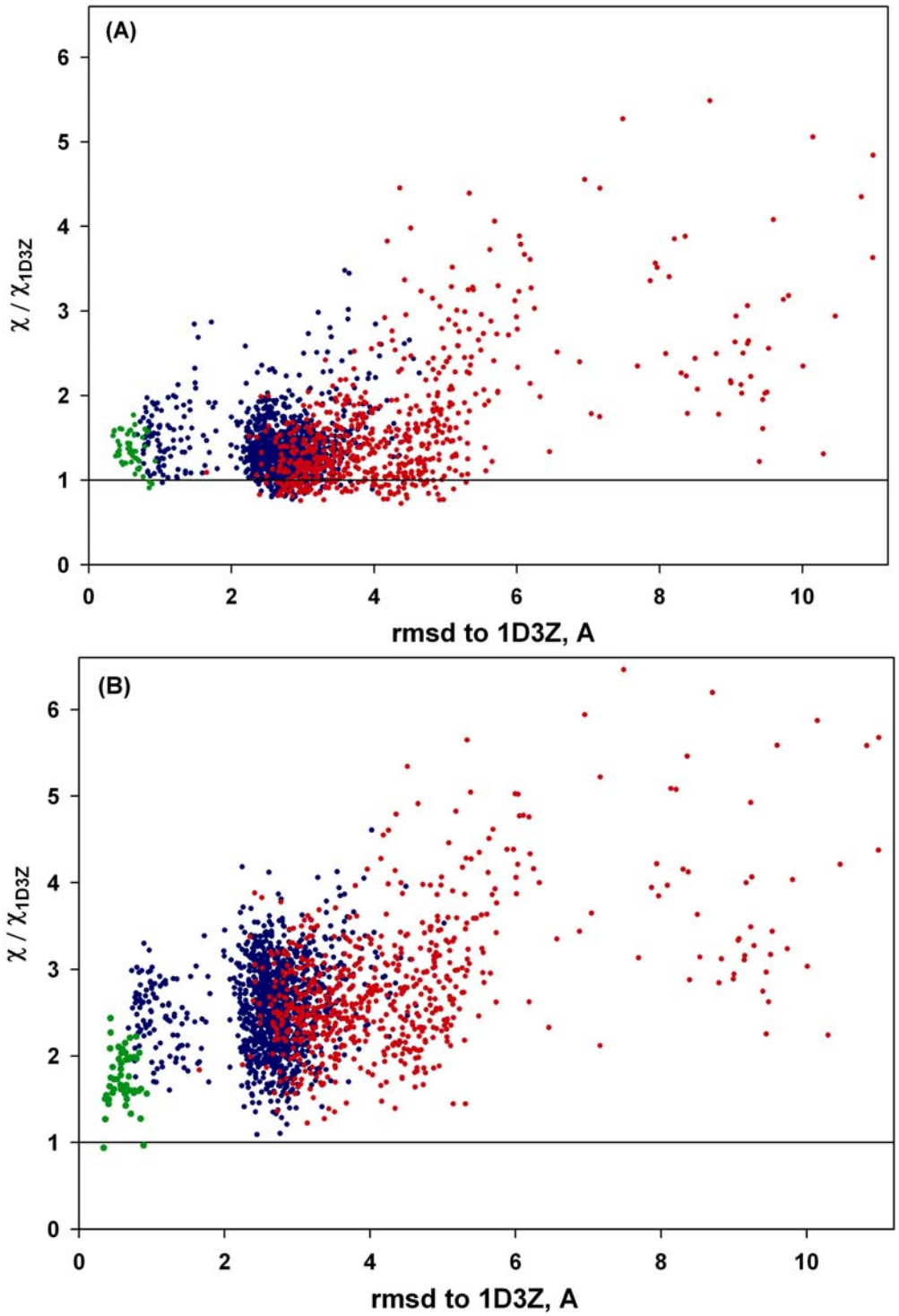


**Figure S2. SAXS data fit for ensembles of ubiquitin models.** Fits for the Rosetta (red), CS-Rosetta (blue), and the 53-member ensemble of crystal structures of full-length ubiquitin (see legend to Fig. S1) are shown. Panel (A) shows Crysol results and panel (B) shows AXES results. Crysol fits were performed using version 2.6 in the batch mode with no explicit hydrogens, a maximum order of harmonics of 50, a Fibonacci grid order of 18, and default settings for the ranges of all other parameters. AXES fits were performed averaging over 10 water boxes for each structure, a Fibonacci grid order of 16, and default settings for all other parameters.

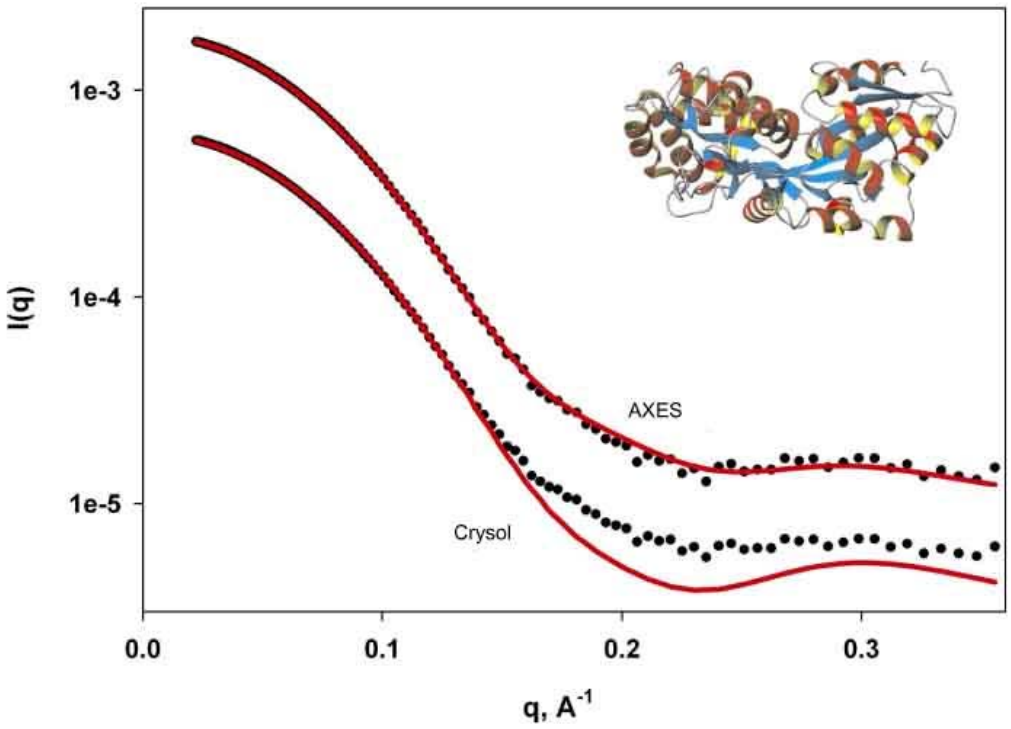


**Figure S3.** Crysol and AXES fits for the ligand-bound 41-kDa Maltose Binding Protein (PDB entry 3MBP)[13]. Fits by AXES ($\chi$=1.00) and Crysol ($\chi$=4.26) are offset for clarity. Experimental data were collected at BeamLine 18-ID (BioCAT), Advanced Photon Source, Argonne National Laboratory, on a 4.2 mg/mL sample. The Crysol fit was performed using program version 2.5, a maximum order of harmonics of 50, a Fibonacci grid order of 18, and default settings for the ranges of all other parameters. AXES fits were performed averaging over 50 water boxes for each structure, using 16 for the Fibonacci grid order, and default settings for all other parameters.

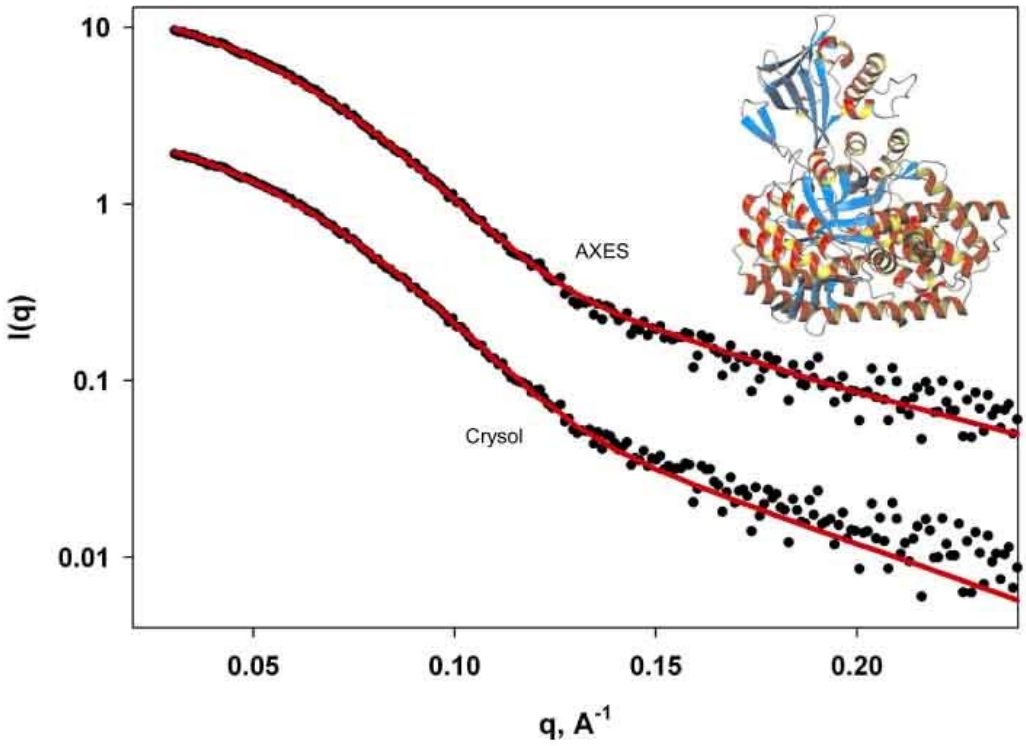


**Figure S4.** Crysol and AXES fits for the 82 kDa Malate Synthase G (PDB entry 1D8C)[14]. Fits by AXES ($\chi$=0.885) and Crysol ($\chi$=0.974) are offset for clarity. Experimental data were collected at BeamLine 4-2 of the Stanford Synchrotron Research Laboratory, on a 14 mg/mL sample. The Crysol fit was performed using program version 2.5, with no explicit hydrogens, a maximum order of harmonics of 50, a Fibonacci grid order of 18, and default settings for the ranges of all other parameters. AXES fits were performed averaging over 50 water boxes for each structure, using 16 for the Fibonacci grid order, and default settings for all other parameters.

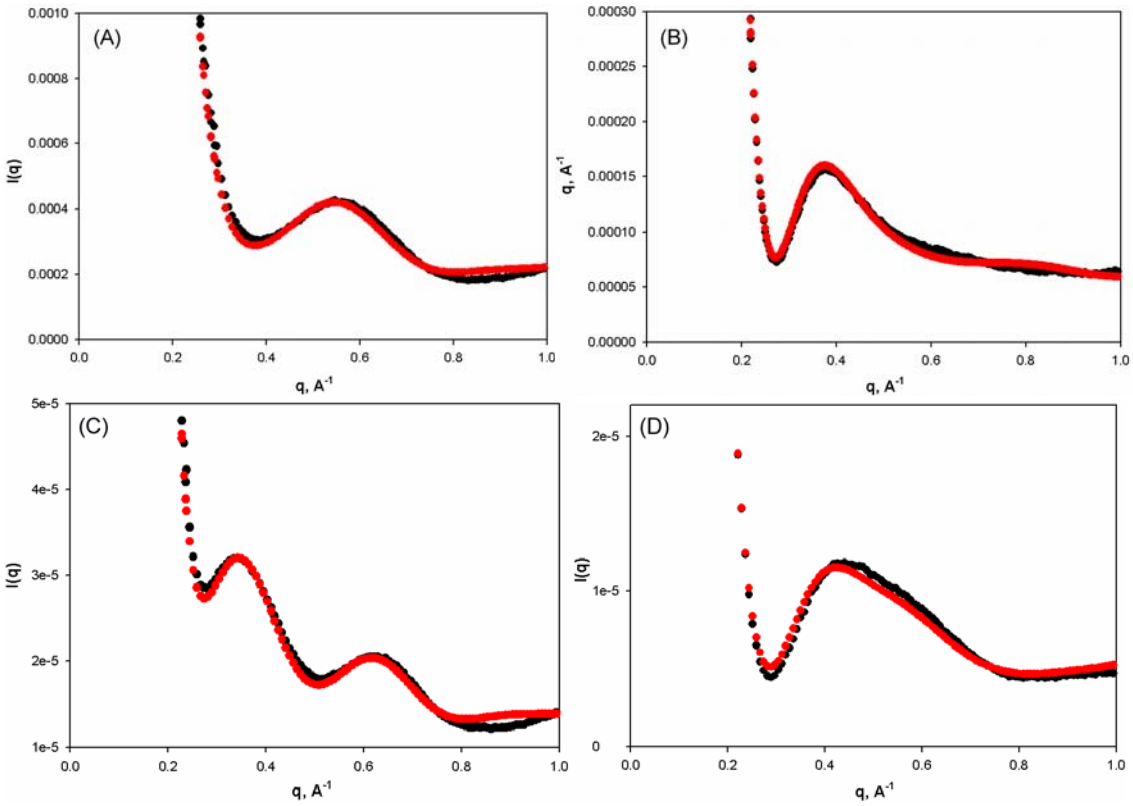


**Figure S5.** Expanded views of AXES fits displayed in Figure 1B, main text, displayed on a linear scale. Panel (A): GB3, panel (B): cytochrome C, panel (C): lysozyme, panel (D): ubiquitin.

**References**


(1) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910-8922.
(2) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089-10092.
(3) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781-1802.
(4) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441-&.
(5) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087-1092.
(6) Sorenson, J. M.; Hura, G.; Glaeser, R. M.; Head-Gordon, T. *J. Chem. Phys.* **2000**, *113*, 9149-9161.
(7) Ulmer, T. S.; Ramirez, B. E.; Delaglio, F.; Bax, A. *J. Am. Chem. Soc.* **2003**, *125*, 9179-9191.
(8) Gasteiger, E.; Gattiker, A.; Hoogland, C.; Ivanyi, I.; Appel, R. D.; Bairoch, A. *Nucleic Acids Res.* **2003**, *31*, 3784-3788.
(9) Putnam, C. D.; Hammel, M.; Hura, G. L.; Tainer, J. A. *Q. Rev. Biophys.* **2007**, *40*, 191-285.
(10) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128-132.
(11) Lange, O. F.; Lakomek, N. A.; Fares, C.; Schroder, G. F.; Walter, K. F. A.; Becker, S.; Meiler, J.; Grubmuller, H.; Griesinger, C.; de Groot, B. L. *Science* **2008**, *320*, 1471-1475.
(12) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836-6837.
(13) Quiocho, F. A.; Spurlino, J. C.; Rodseth, L. E. *Structure* **1997**, *5*, 997-1015.
(14) Howard, B. R.; Endrizzi, J. A.; Remington, S. J. *Biochemistry* **2000**, *39*, 3156-3168.