Journal of
Applied
Crystallography
ISSN 0021-8898

Received 16 August 2007
Accepted 20 February 2008

Small-angle neutron scattering and the errors in protein structures that arise from uncorrected background and intermolecular interactions

Kenneth A. Rubinson,[a,b]* Christopher Stanley[a,c] and Susan Krueger[a]*

[a]NIST Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA, [b]Department of Biochemistry and Molecular Biology, Wright State University, Dayton, OH 45435, USA, and [c]Laboratory of Physical and Structural Biology, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, USA. Correspondence e-mail: rubinson@nist.gov, susan.krueger@nist.gov

Small-angle neutron scattering (SANS) provides a unique method to probe soft matter in the 10–100 nm length scale in solutions. In order to determine the shape and size of biological macromolecular structures correctly with SANS, a background-subtracted, undistorted scattering curve must be measured, and the required accuracy and precision is especially needed at the short-length-scale limit. A true scattering curve is also needed to discern whether intermolecular interactions are present, which also are probed in the SANS experiment. This article shows how to detect intermolecular interactions so that subsequent structure modeling can be performed using only data that do not contain such contributions. It is also shown how control of many factors can lead to an accurate baseline, or background, correction for scattering from proteins, especially to account for proton incoherent scattering. Failure to make this background correction properly from proteins, polymers, nucleic acids and lipids can result in incorrect values for the calculated shapes and sizes of the molecules as well as the derived magnitudes of the intermolecular interactions.

1. Introduction

Measurements on solutions of biological macromolecules (and other 'soft matter') by small-angle neutron scattering (SANS) are becoming ever more routine. Most commonly, the scattering is used to determine radii of gyration of the scattering entities (Guinier, 1939; Porod, 1982), their molecular weights from extrapolation of the scattering to its intensity at zero angle (Jacrot & Zaccai, 1981) and their shapes.

The scattering experiment consists of measuring the amount of neutron scattering as a function of $q = (4\pi/\lambda)\sin\theta$, where λ is the neutron wavelength and 2θ is the scattering angle measured from the axis of the incoming neutron beam. The dependence of the scattering intensity, $I(q)$, on q can be separated into a factor describing scattering from individual particles, the form factor or shape factor $P(q)$, and a factor describing scattering from the sum of pairs of particles, the interparticle structure factor $S(q)$. From modeling $I(q)$, molecular sizes and shapes can be approximated from $P(q)$ (*e.g.* Kline, 2006).

In addition, H–D isotopic substitutions in the solvent can be used to dissect the structures of selected parts of larger molecular constructs or to ascertain individual structures of different molecules that coexist in a solution. This is called contrast variation, where the contrast is the difference in the strength of the scattering interaction of the solute molecules compared with that of the solvent. The contrast variation technique can separate, for example, scattering of DNA from that of proteins in the same solution.

Neutron scattering found from a SANS experiment on solutions of a polymer or protein can be classified into two general types: incoherent and coherent (Squires, 1996). The incoherent scattering arises from the neutrons being scattered by single nuclei. The incoming neutrons are scattered randomly in direction, and the magnitude of this scattering depends on the identity of the nuclei and their concentrations. Hydrogen scatters incoherently about 40 times as much as deuterium. On the other hand, coherent scattering arises from two nuclei separated by a distance of the order of the neutron de Broglie wavelength. The structural information we seek comes from interpreting the coherent scattering as a function of q. The incoherent scattering is a source of background that must be correctly subtracted from the total scattering in order to obtain the correct coherent scattering contribution from the molecule, which can then be used to determine accurate structural parameters.

2. Statement of the problem

Incorrect subtraction of the baseline (or background) scattering contributes to a number of problems, including misinterpreting molecular sizes and shapes. Also less accurate are molecular weights calculated from the extrapolation to $I(0)$, the scattering intensity at $q = 0$ (scattering angle of 0°), by the method of Jacrot & Zaccai (1981), and the radius of gyration, R_g, calculated by the method of Guinier (Guinier, 1939; Porod, 1982). Even when the correct baseline is subtracted from the measured SANS data, the weight-average molecular weight, M_w, and geometric parameters can still be inaccurate because of the presence of intermolecular interactions that perturb the scattering form factor.

Four major points to consider in order to obtain reliable structural information for biomolecules in solution from SANS intensities, $I(q)$, are:

(1) Baselines must be determined experimentally, because of nonlinearities in hydrogen incoherent scattering contributions. This is especially true for molecules in mixed H–D solvents, as used for contrast variation studies.

(2) Contrast match points should be measured experimentally, since the point where the solute and solvent scatter equally is unknown because of H–D exchange of hydrogens in the molecule with solvent deuterium. This uncertainty in contrast also affects the accuracy of calculated values of M_w and R_g.

(3) The effect of intermolecular interactions on $I(q)$ must be determined experimentally. Model structures obtained assuming only a form factor can be confounded by molecule–molecule interactions at remarkably low concentrations.

(4) When comparing data with model SANS curves calculated from high-resolution X-ray or NMR structures, adjusting a single parameter representing the baseline level often is as good as using a multi-parameter fit. In parallel, parameters used to fit low-resolution model structures to SANS data are coupled to each other and can compensate for changes in each other to produce wide variations in the possible models that can fit the SANS data equally well.

These four points are expanded in order below in similarly numbered subsections of §4.

3. Materials and methods[1]

3.1. Sample preparation

The light water used was deionized, with a resistance of 18.2 MΩ (Millipore, Billerica, MA). D_2O (99.9 mol% D, nominally 100% D_2O) was obtained from Cambridge Isotope Laboratories Inc. (Andover, MA). Samples of various volume-to-volume (*v*/*v*) mixtures of D_2O and H_2O were held in 1.000 mm (1 mm) or 2.000 mm (2 mm) path length cylindrical quartz cuvettes (Hellma USA, Plainview, NY).

Poly(ethylene glycol) of $M_w = 400$ g mol^{-1} (Da) (PEG 400) was purchased from Hampton Research (Aliso Viejo, CA). PEG 400 solutions were made from 50% weight-to-volume (*w*/*v*) stock solutions in D_2O. If the stock solution was not between pH 6 and pH 8, it was brought into that range by adding DCl or NaOD in D_2O. Stock solutions of salts and buffers were added to produce test samples with the appropriate PEG concentration, together with 10 m*M* HEPES (1:1 acid:Na salt, pK_a in H_2O 7.55), 100 m*M* ammonium sulfate and 0.1%(*w*/*v*) NaN_3. The final solution was made at least 24 h in advance to allow equilibration.

Lysozyme from chicken egg white (3× crystallized, dialyzed and lyophilized) was purchased from Sigma (St Louis, MO) and used without any further purification. Lysozyme solutions for neutron scattering experiments were prepared at nominal concentrations of 5 mg ml^{-1} (0.5% *w*/*v*), 10 mg ml^{-1} and 20 mg ml^{-1} in 20 m*M* potassium acetate (pH 5), 100 m*M* KCl, D_2O buffer. The protein concentration for the nominal 5 mg ml^{-1} sample was 4.7 mg ml^{-1} as determined spectrophotometrically by the absorbance at 281.5 nm wavelength ($A_{281.5}$ = 0.38 for 1 mg ml^{-1} for a 1 cm path length). The samples were loaded into 2 mm path length quartz cuvettes for the SANS measurements.

3.2. SANS measurements

SANS measurements were performed on the NG7 and NG3 30 m SANS instruments at the NIST Center for Neutron Research (NCNR) in Gaithersburg, MD (Glinka *et al.*, 1998). Scattered neutrons were detected with a 64 × 64 cm two-dimensional position sensitive detector with 128 × 128 pixels and 0.5 cm resolution per pixel. Data reduction was accomplished using *Igor Pro* software (WaveMetrics, Lake Oswego, OR) with SANS macros developed at the NCNR (Kline, 2006). Raw counts were normalized to a common monitor count and corrected for empty cell counts, ambient room background counts and non-uniform detector response. Data were placed on an absolute scale by normalizing the scattering intensity to the incident beam flux. Finally, the data were radially averaged to produce the scattering intensity $I(q)$ to plot as $I(q)$ *versus* q curves.

All water samples were measured at 295 K using a sample-to-detector distance of 1.5 m with a detector offset of 20 cm. SANS data for the H_2O–D_2O mixtures were obtained using λ values between 5.2 and 5.5 Å, with $\Delta\lambda/\lambda$ values between 0.11 and 0.15. Some measurements were also made using λ = 8 Å with $\Delta\lambda/\lambda$ = 0.15. Data were collected for 5 min for each sample. After data reduction, about 100 data points between q = 0.0170 and 0.1407 Å^{-1} of the non-sloping line were averaged. The relative standard deviations of the average $I(q)$ obtained this way, $\langle I(q)\rangle$, were less than 5%. Then, $\langle I(q)\rangle$ of the 100% D_2O sample was subtracted from the $\langle I(q)\rangle$ values of each mixture, which sets the 100% D_2O samples as $I(q) \equiv 0$.

The PEG 400 experiments were performed using λ = 5.2 Å with $\Delta\lambda/\lambda$ = 0.15. If evidence for gas bubble scattering was present, the solutions were degassed in the liquid form under vacuum. The lysozyme samples were measured using λ = 5.5 Å with $\Delta\lambda/\lambda$ = 0.11.

[1] Certain trade names and company products are identified in order to specify adequately the procedure. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products are necessarily the best for the purpose.

Table 1
Scattering length densities, ρ, of compounds in units of $10^{10}\,\mathrm{cm}^{-2} \equiv 10^{-6}\,\mathrm{\AA}^{-2}$.

% D_2O	Water	Proteins	DNA	PEG 400
0	−0.56	1.77	3.68	0.64
100	6.38	3.09	4.76	0.97

3.3. SANS data analysis

3.3.1. Radius of gyration calculation. From geometry, the radius of gyration, R_g, is the second moment of rotation about the centroid of the scattering particle. $R_g = (1/V)\int_V \rho(r)\, r^2\, \mathrm{d}V$, where V is the particle volume and ρ is its scattering length density, defined as the sum of the scattering lengths of all the individual atoms in the particle divided by the particle volume. Table 1 shows some typical values of ρ for compounds of interest.

R_g was calculated from the SANS data using the Guinier (1939) approximation, $I(q) \simeq P(q) \simeq I(0)\exp(-q^2 R_g{}^2/3)$, which is valid when $qR_g < 1$. The data were plotted as $\ln[I(q)]$ *versus* q^2, and a straight line was found over a range of q that includes $qR_g \simeq 1$. The slope of the straight line is

$$\mathrm{d}\ln I(q)/\mathrm{d}q^2 = -R_g^2/3, \tag{1}$$

from which R_g was found.

3.3.2. Molecular weight calculation. The molecular weight was found from $I(0)$ using the equation

$$I(0) = n(\Delta\rho)^2 V^2, \tag{2}$$

where n is the number density of particles in cm^{-3} and $\Delta\rho$ is the contrast in units of cm^{-2}. $\Delta\rho = \rho - \rho_s$, where ρ_s is the scattering length density of the solvent. V is in units of cm^3. Equation (2) assumes that the data are on an absolute scale, usually in units of cm^{-1}, that the background has been subtracted completely, that the solute is monodisperse and that intermolecular interactions are effectively absent.

The molecular weight of the scatterers was found from $I(0)$, without requiring structure information (Jacrot & Zaccai, 1981), by making the following substitutions in equation (2):

$$n = cN_A/M_w, \quad V = \bar{v}M_w/N_A, \tag{3}$$

where c is the particle concentration in $\mathrm{g\,cm}^{-3}$, N_A is Avogadro's number and $\bar{v}$ is the average partial molar volume of the solid molecule in $\mathrm{cm}^3\,\mathrm{g}^{-1}$. For PEGs, this value is $0.89\,\mathrm{cm}^3\,\mathrm{g}^{-1}$. Protein values are found in the range 0.70–$0.76\,\mathrm{cm}^3\,\mathrm{g}^{-1}$, with a large number clustered around $0.74\,\mathrm{cm}^3\,\mathrm{g}^{-1}$ (Perkins, 1986). Together, the substitutions in equation (3) produce

$$I(0) = (cM_w/N_A)\,(\Delta\rho)^2\,\bar{v}^2, \tag{4}$$

which, with the measured value for c and the calculated values for $\Delta\rho$ and $\bar{v}$, was then used to calculate M_w from the fitted value of $I(0)$.

3.3.3. Nonparametric calculation for $S(q)$ from scattering curves. In practice, the scattering that is measured from biological macromolecules in solution includes contributions from the form factor, $P(q)$, and the interparticle structure factor, $S(q)$, as well as background scattering from solvents, buffers and cuvettes. This background scattering must be subtracted from the total measured scattering in order to obtain the scattering intensity from the macromolecules alone. The contributions to the measured scattering intensity are related in equation (5):

$$I(q) = (\partial\sigma/\partial\Omega) = n\,V^2(\Delta\rho)^2 P(q)\,S(q) + B(q), \tag{5}$$

where $\partial\sigma/\partial\Omega$ is the differential cross section in cm^{-1} and $B(q)$ is the total background signal from the solvent, buffer, cuvette and solutes, which includes incoherent scattering from hydrogen.

The form factor $P(q)$ was separated from the interparticle structure factor $S(q)$ by obtaining the scattering curves at two different concentrations; call them 1 and 2. However, for scatterers that do not change shape with concentration (proteins are generally good examples), only the terms $S(q)$, n and $I(q)$ vary. If the unchanging factors are given the subscript p for particle,

$$\begin{aligned} I(q)_1 &= n_1\,V_p^2\,(\Delta\rho)_p^2\,P(q)_p\,S(q)_1 + B(q)_1 \quad \text{and} \\ I(q)_2 &= n_2\,V_p^2\,(\Delta\rho)_p^2\,P(q)_p\,S(q)_2 + B(q)_2. \end{aligned} \tag{6}$$

If the backgrounds $B(q)_1$ and $B(q)_2$ can be subtracted exactly, they may be ignored, and the following ratio is true:

$$\frac{I(q)_1}{I(q)_2} = \frac{n_1}{n_2}\,\frac{S(q)_1}{S(q)_2}. \tag{7}$$

We assume that, at the lower concentration, interparticle interactions are negligible, and at the higher concentration, some interaction may occur. At the lower, noninteracting, concentration – call it 1 – the value of $S(q)_1$ is unity at all q. This means that, as pairs, the particles do not change the scattering. As a result, for the higher concentration sample, $S(q)_2 = S(q)_{\mathrm{interacting}}$ was found by using the scattering data directly without needing a structure model for the scattering particles.

$$S(q)_2 = \frac{n_{\mathrm{noninteracting}}}{n_{\mathrm{interacting}}}\,\frac{I(q)_{\mathrm{interacting}}}{I(q)_{\mathrm{noninteracting}}}. \tag{8}$$

Any measure of concentration can be used with equation (8), and molarity or volume fraction are convenient.

As shown by Hayter & Penfold (1983), this separation of $P(q)$ and $S(q)$, which allows equation (8) to be employed, strictly holds only for homogeneous monodisperse spheres in solution. However, it has been found to work for solutes that are not spheres and not strictly monodisperse. We assume that equations (4), (5) and (8) hold exactly rather than as approximations. This is standard practice for analyzing the data from SANS of proteins.

3.4. SANS *I*(*q*) modeling

Calculated SANS $I(q)$ curves were derived from high-resolution X-ray crystal structures using the programs *CRYSON* (Svergun *et al.*, 1998) and *XTAL2SAS* (Krueger *et al.*, 1998). *CRYSON* is a neutron-specific version of *CRYSOL*

(Svergun *et al.*, 1995) that calculates model SANS intensities using spherical harmonics. *XTAL2SAS*, on the other hand, uses a real space approach first described by Heidorn & Trewhella (1988). In *XTAL2SAS*, each amino acid residue in the protein is represented as a sphere of an appropriate scattering length density and size. The spheres are randomly filled with points using a Monte Carlo method (Hansen, 1990). All possible pairs of points are then summed to form a histogram of distances in the molecule, and the model SANS curves are calculated by a Fourier transform of this distance distribution function.

3.5. Working with previously published data

SANS data taken from the literature were digitized from the published graphs using the software *UnScanit* (Silk Scientific, Orum, UT). SANS data for horse heart cytochrome c were taken from Fig. 5 in the paper by Wu & Chen (1987), in which the protein was in a 100 m*M* sodium acetate buffer, pH 6.8, with 0.02% NaN_3. The $S(q)$ graphs for cytochrome c were calculated using equation (8) after fitting the digitized $I(q)$ data for 0.45, 0.91 and 1.81%(w/v) solutions with closely fitting Gaussian functions and then assuming the 0.45% solution to be non-interacting. All curve fitting was performed with *TableCurve 2D* (Systat, San Jose, CA).

4. Discussion

Four major points to consider to obtain properly corrected scattering intensities were listed in §2. We now discuss these points in numerical order below, with the first decimal having the corresponding number.

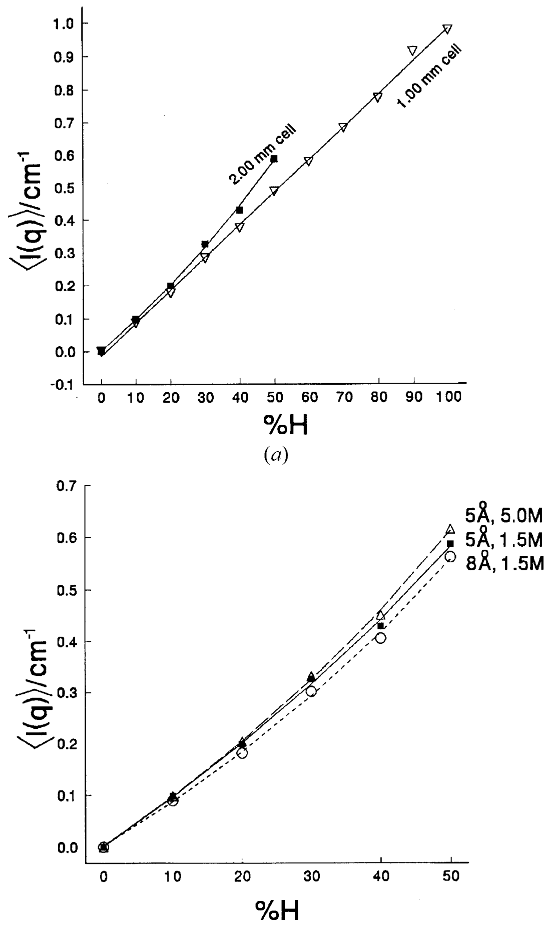
4.1. Determining correct baselines

4.1.1. Baseline corrections for incoherent scattering. The undesired incoherent scattering that arises from the hydrogen in the buffer must be measured as accurately as possible under the same conditions used for measuring the sample in order to subtract it completely. Nonlinearities in proton incoherent scattering mean that measurements made under one set of conditions cannot be extrapolated easily to other conditions. These include different path lengths or different amounts of hydrogen (H) and deuterium (D) in the solvent. The reasons for this nonlinearity include a large number of different phenomena that depend on the significantly different scattering properties of H and D. For example, the background incoherent scattering depends on the sample geometry (Carsughi *et al.*, 2000; May *et al.*, 1982; Shibayama *et al.*, 2005), and on the amount of multiple scattering, which is a function of wavelength and isotope concentration (Shibayama *et al.*, 2005), as well as on the detector's sensitivity to the energies of the scattered neutrons. These energies also depend on the sample's composition and geometry (Ghosh & Rennie, 1999).

To illustrate the nonlinearity in proton incoherent scattering, the scattering from water mixtures with various ratios of H and D were measured using both 1 and 2 mm path length cuvettes. The 1 mm path length data shown in Fig. 1(a) were fitted with a straight line. However, a straight line would not fit the 2 mm data, so the 2 mm data in both Figs. 1(a) and 1(b) were fitted with the simple exponential

$$\langle I(q)\rangle = a + b\exp(\%\mathrm{H}/k), \tag{9}$$

where %H is the mole percent hydrogen content, and a, b and k are arithmetic fitting constants. Owing to the above-mentioned complexity of the phenomenological contributions, the fitting variables have no clear physical meaning. As can be seen in Fig. 1, systematic variations in the background arise from changes in sample path length, detector position and neutron wavelength even when a common zero is chosen for 100% D_2O. Not shown are curves from a different SANS beamline at the NCNR, NG7; those curves exhibit systematic differences from the data from beamline NG3 seen here. It can be concluded from these results that corrections for incoherent background scattering cannot be transferred to background measurements made under different experimental conditions or to a different instrument.



(a)

(b)

Figure 1
Data of the values of $\langle I(q)\rangle$ [equation (9)] for scattering of water in 1 and 2 mm path length fused silica cells as a function of the mole percent of hydrogen (%H), measured on the NG3 SANS instrument at the NCNR. (a) Data for 1 and 2 mm cells with the same neutron wavelength (5 Å) and detector distance (1.5 m). (b) Data for 2 mm cells with different sets of wavelength and detector distances, as marked near the curves.

Besides the main incoherent scattering contribution from the buffer, there is also an incoherent contribution from the protons in the solute. The volume fraction of hydrogens may change with solute concentration and with it the incoherent scattering from the solutions. Therefore, if the concentration of the solute is high enough, it may not be sufficient merely to subtract the scattering of the buffer from that of the solution in order to correct for the background scattering. The 'excess' incoherent scattering from the solute must also be subtracted from the measured scattering intensity in order to obtain an $I(q)$ curve that is free of incoherent scattering. If the solute is in a buffer deuterated to some extent, a buffer with an H–D ratio that compensates for the proton content in the solute could possibly be used to obtain a more accurate baseline subtraction.

The importance of making as precise a baseline subtraction as possible cannot be overstated. Even when the most careful buffer subtraction is made, it is sometimes difficult to determine whether any excess incoherent scattering from the solute is still present in the data. However, if an effort is made to make the most exact baseline correction possible, then the corrected $I(q)$ will be as accurate as possible and the structural parameters determined from the data will be more trustworthy.

4.1.2. Contribution of incoherent scattering to R_g.

To demonstrate the effect of the correction for incoherent scattering, we measured a number of samples of PEG, which has the simplifying advantage over proteins that it does not exchange any of the chain hydrogens with the D_2O solvent. The correction to be made requires a knowledge of the number density of hydrogens in the solute and in the solvent. However, the mass density of hydrogen is more convenient to use. For example, in water, with a density of 1.00, hydrogen makes up 2/18 of the mass, and so the relative mass density is 0.11. The calculation is similar for PEG 400 [for formula weight 414, the formula is $HO(OCH_2CH_2)_9OH$]; the specific gravity of the neat liquid is 1.12, and the relative mass density for the hydrogen is 38/414 = 0.092, so the mass density is 0.092 × 1.12 = 0.10. The hydrogen densities are, conveniently, essentially the same for water and PEG. The volume fractions, then, are quite good measures of the H–D ratio, and we substitute the volume fraction for the number density.

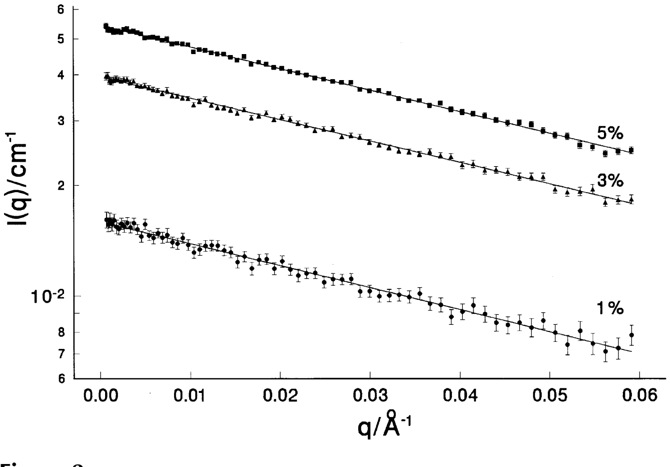


Figure 2
Guinier plots of the scattering data from 1, 3 and 5%(w/v) D_2O solutions of PEG 400 corrected for solvent background and proton incoherent scattering.

Table 2
R_g of PEG 400 from Guinier plots of SANS data corrected and uncorrected for proton incoherent scattering.

	Corrected		Uncorrected	
Concentration %(w/v)	R_g (Å)	Linear plot range	R_g (Å)	Linear plot range
1	6.43 (7)	$0.17 < qR_g < 1.62$	5.04 (8)	$0.37 < qR_g < 1.10$
3	6.34 (4)	$0.33 < qR_g < 1.39$	4.16 (4)	$0.74 < qR_g < 1.21$
5	6.34 (3)	$0.33 < qR_g < 1.39$	3.98 (3)	$0.73 < qR_g < 1.21$

Because of the increased incoherent scattering with added ^{1}H–PEG in D_2O, the magnitude of a measured $I(q)$ curve is greater than it otherwise would be. An example of the trend is shown in Fig. 2, the Guinier plots of data of PEG 400 in D_2O at three different concentrations. The resulting R_g values appear in Table 2. The contribution from extra incoherent scattering needs to be subtracted. Its value is independent of q, so for each condition a single number needs to be subtracted from all the data points. When this correction is made for the 1, 3 and 5%(w/v) solutions of PEG 400, the calculated value of R_g is constant, as shown in Table 2. PEG 400 was chosen because the results are highly sensitive to the baseline position at high q since, because of its small size, the molecules are expected to be scattering up to and past the high-q cutoff of the data. However, the radius of gyration calculated is sensitive to the baseline-corrected scattering, and the more precisely the correction can be made, the more consistent the result. For example, as listed in Table 2, when the incoherent scattering correction is neglected, the molecules appear to shrink with increasing concentration. Similar misinterpretations are possible for proteins, and these are discussed further below.

Since the incoherent scattering background may be nonlinear, corrections for solutes dissolved in pure D_2O differ in magnitude from the corrections made for H_2O–D_2O solutions. For example, when PEG is in pure D_2O, the corrections made are for 1, 3 and 5% volume fractions of hydrogen. However, if the PEGs are added to a 50%(v/v) D_2O solvent, the applicable correction requires subtracting the difference of the incoherent scattering between light water solutions of 51, 53 and 55%(v/v) and the baseline 50% solvent.

4.1.3. Contribution of incoherent scattering to apparent molecular weight.

Not only R_g depends on correcting for incoherent scattering; so does a calculation of M_w for the particles as found from $I(0)$ as defined in equation (2). In practice, the calculated value of M_w is an average of molecular weights if the molecules are not monodisperse. In addition, in practice, there is an unavoidable uncertainty associated with $I(0)$ and c. The accuracy of the absolute scaling of $I(0)$ has a few percent error, and an error of similar magnitude arises in measuring the protein concentration either by UV absorption or other typical assays. As a result, in our experience, $I(0)/c$ typically is accurate to about ±5%.

Table 3
Molecular weights of PEG 400 from extrapolated $I(0)$ corrected and uncorrected for proton incoherent scattering.

Concentration (% *w*/*v*)	M_w (Da) corrected†	M_w (Da) uncorrected†
Extrapolated to 0	(391)	(488)
1	364	465
3	301	419
5	249	374

† Uncertainty of M_w is ±5%.

Equation (4) was used to find the molecular weights of the PEG samples from the extrapolated values of $I(0)$ corrected for proton incoherent scattering. The results are shown in Table 3, where the M_w values obtained with and without the incoherent background correction are compared. The meaning of these results is discussed in more detail in §4.3.

4.2. H–D exchange and hydration water structure

A unique tool of SANS is control of the contrast between solvent and solute by mixing hydrogen and deuterium solvents. For proteins this means primarily H_2O and D_2O. Such control allows the separation of the scattering arising from different scatterers in the same solution. This idea is illustrated in Fig. 3, which shows the scattering length density as it depends on the fraction of D_2O in an aqueous solvent. The scattering length density of PEG 400 matches that of a solvent containing about 15% D_2O. For proteins, the value is about 40% D_2O, and for DNA, the value is about 70% D_2O. These matching concentrations lie where the line graphing $\rho_{\rm water}$ intersects the lines of the ρ values for the three types of solutes. These intersection points are called the match points of the solutes.

Because the contrast associated with the scattering is the difference between the scattering length densities of the molecules and the solvent, any solvent associated with the molecule that has the same weight density as the bulk provides no contrast. As a result, the molecular weight from SANS is often said to be the 'dry' weight.

However, the exact match points of solutes often must be measured, since they cannot be calculated exactly because some fraction of the hydrogen of the solutes exchange with the solvent deuterium. This exchange causes the scattering length densities of the solutes to change with the change in solvent composition. For example, PEG 400 only exchanges the terminal alcoholic hydrogens, which is easy to account for exactly in a scattering length density calculation. Even this exchange can be ignored for higher molecular weight PEG polymers. Proteins can exchange backbone amide hydrogens and some side-chain hydrogens, and the fraction of H–D exchanged depends on the amino acid content and the structure (Efimova *et al.*, 2005; Harrison *et al.*, 1988; Perkins, 1986). However, since the amount of H–D exchange in a protein is less certain, so is the calculated contrast variation, which in turn makes estimated molecular weights and radii of gyration less exact.

A general rule for proteins is that 10% of the volume fraction of hydrogen of the backbone do not exchange, but all side-chain hydrogens exchange (Perkins, 1986). Some details of these exchanges have been investigated by mass spectrometry (Efimova *et al.*, 2005) and by neutron crystallography (Harrison *et al.*, 1988). Overall, the scattering length densities of proteins change about a fifth as fast with H–D substitution as that of the suronding water (see Table 1). Match points for proteins range from 39% D_2O to 45% D_2O, with many clustered in the 41–43% range (Perkins, 1986).

Some of the background subtraction problems that arise when using mixed H–D solvents can be mitigated by dialyzing the proteins into the appropriate solvent and using the dialysate for the baseline determination. Then, the final proton content of the solvent is the same for both the sample and the dialysate blank. However, an uncertainty in the total proton concentration in the solute remains, and the correction to the background must still be discovered experimentally as described in §4.1.

4.2.1. The question of contrast for hydration water. Details of protein H–D exchange also have an effect on the contrast that might be seen in the hydration layer of proteins. In order to delineate the hydration layer, it must have a contrast differing from that of the solvent and from the protein, and it must be a distinct layer with a clear boundary. The known chemistry and structure of protein surfaces strongly suggest that neither having a distinct layer nor showing contrast with the solvent are supportable. The two confounding factors are that the surface is a 'blur' rather than a clear layer, and the water density is unlikely on average to be different from the surrounding solvent. Details of the applicable chemistry follow.

First, the surface of a protein provides a blurred transition for a number of reasons. By blurred, we mean that there is heterogeneity in the contrast over distances of about 5 Å. This heterogeneity arises from a number of sources. There is the interplay between freely mobile side chains of the protein and the hydration water molecules; the water molecules may be between the side chain and the main body of the protein as

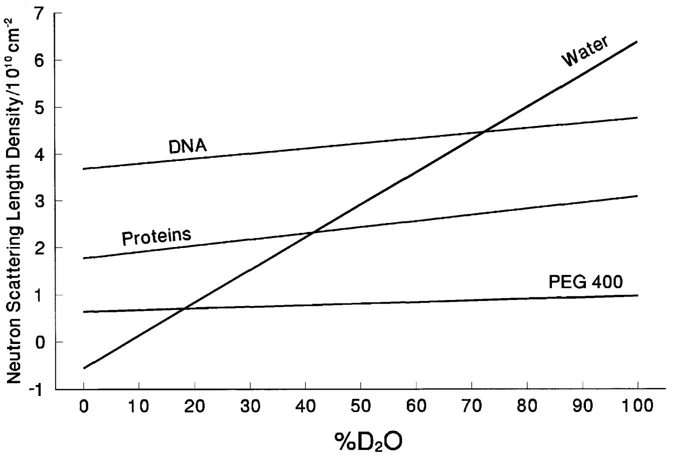


Figure 3
The values of neutron scattering length density for DNA, proteins, PEG 400 and water as a volume percentage of D_2O in the water solvent.

well as having the side chain and water positions reversed. The SANS scattering consists of the average of a series of 'flash' images (in q space) of all the accessible structures. The measured structure provides only the average of the water molecules and side chains. As a result, water/side-chain structures are uncorrelated with the bulk of the protein and contribute proportionally less to the scattering (Levitt & Park, 1993). Finally, as will be shown in §4.4, SANS is not sensitive to the details of how the contrast changes at the boundary between a scatterer and the solvent.

Further blurring occurs from local chemistries. At the hydrophilic and charged sites of the protein surface, hydrogen-bonded water molecules, which have a relatively fixed structure consisting of a linear alignment of the hydrogen between two heavier atoms, are heterogeneous in their directions from the surface. After all, the atoms of the protein to which the water molecule hydrogen bonds, say the backbone amide nitrogens, are relatively randomly oriented relative to the surface normal. Therefore, the hydrogen-bonded water molecule is expected to be orientationally disordered. Furthermore, the surface is uneven on the water scale, and not only does this provide a blurred edge, but water may be excluded from some recesses and fitted into others (Edison *et al.*, 1995; Levitt & Park, 1993). Any exclusion decreases the expected average density. As a result, even if charged or neutral groups are assumed to provide a denser hydrogen volume fraction nearby (which is not obviously observable in coherent scattering because of the orientational disorder), the higher density could be countered by exclusions from recesses.

Even more disorder and variability in the water density is expected because of the heterogeneity of the surface hydrophilicity and hydrophobicity. As well documented by Chalikian *et al.* (1996), a broad range of water-soluble proteins have about half of their surfaces covered by hydrophilic or charged groups and half covered by neutral groups, some of which are hydrophobic. As has been shown by neutron reflectometry (Schwendel *et al.*, 2003), the density of water at a flat hydrophobic surface is lower than bulk; the water suffers dilation of about 10% (Ball, 2003; Jensen *et al.*, 2003). A similar effect could be expected at the part of the surface that is hydrophobic, which would more than counter any possible increased density caused by electrostriction on the relatively few charged areas of the surface (Perkins, 1986). Areas that are hydrophilic, may, as Jensen *et al.* (2003) found from X-ray reflectometry at extended hydrophilic surfaces, have a density of water equal to that in the bulk.

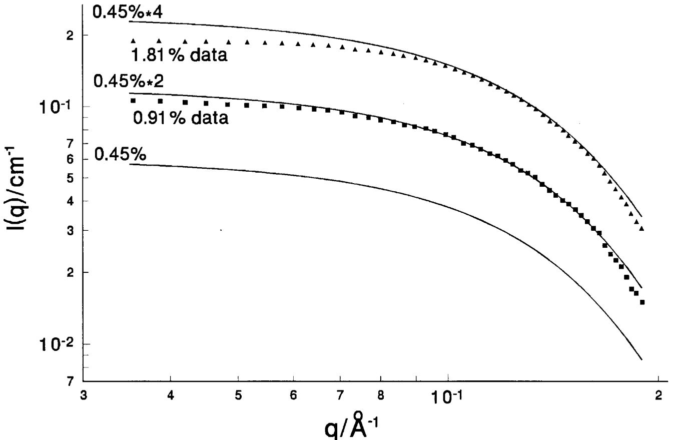


Figure 4
Data of Wu & Chen (1987) for SANS of cytochrome c showing the best fit Gaussian function curve for the 0.45%(w/v) solution (lowest, solid line) and the measured SANS data for the 0.91% (closed squares) and 1.81% (closed triangles) solutions. The two upper solid curves are twice and four times the 0.45% curve.

In other words, the descriptive chemistry of surfaces suggests that water at the surface of a protein is electrostricted at the relatively few charged groups, equal to the bulk at hydrophilic regions, and dilated over the fraction of the surface that is hydrophobic. Not only is the density of the surface water expected to be equal to or less than the bulk, the heterogeneity in the density of the hydration layer over the surface will also contribute to blurring, which further limits the likelihood of observing an effect on SANS from it.

Another factor that could come into play is H–D preferential partitioning; hydrogen-bonding sites might partition one isotope preferentially as well as change the stability of the protein. This would create another mechanism for blurring and cause uncontrollable changes in local scattering length densities. However, major amounts of such fractionation do not appear to occur for proteins and apparently can be ignored (Edison *et al.*, 1995; Schowen & Schowen, 1982).

4.3. Effects of intermolecular interactions

Even when the correct baseline is subtracted from SANS data, in order for the molecular weight and geometric parameters to be precise, intermolecular interactions cannot perturb the form factor. For example, as shown in Table 3, the apparent molecular weight of PEG 400 appears to change with concentration, which is an indication of intermolecular interactions. The calculated masses are inversely and linearly proportional to concentration, and the mass extrapolated to zero concentration can be found easily. The limiting values with and without the correction for incoherent scattering are shown in Table 3. The corrected value is far closer to the nominal molecular weight.

PEG 400 at 1% concentration is 25 mM, while a small protein like cytochrome c with its molecular weight of 12.3 kDa at 1% is only 0.8 mM; its molar concentration is 30-fold more dilute. Nevertheless, intermolecular interactions are clearly perturbing the form factor, as can be seen in Fig. 4. Here, the scattering from the 0.91 and 1.81% solutions are compared with, respectively, two times and four times the scattering from a 0.45% solution. The scattering does not scale linearly, which indicates a perturbation. The perturbation may be chemical, where some monomeric material forms clusters that are much larger than the low-q cutoff, or it may be from intermolecular interactions. In other words, if material is removed from the observed q range, the shapes of the scattering curves would be congruent, but smaller than expected, and the data would form parallel curves on this log–log graph.

Intermolecular interactions are indicated here since the curve shapes differ with concentration.

The $S(q)$ functions for these cytochrome c solutions can be estimated using equation (8), and these are shown in Fig. 5. These curves are expected to approach unity at high q, but at values far outside the data cutoff. If some intermolecular interactions are in fact present even at the lowest concentration measured here, then the $I(q)$ curves would be smaller than for truly non-interacting solutes, and the low-q portion of the $S(q)$ curves shown in Fig. 5 would then appear larger than they should be. The $S(q)$ curves are then upper limits to the true $S(q)$ curves.

The effects of interparticle interactions are not always seen as clearly as in this example. The degree to which the interactions affect $S(q)$ and, subsequently, the shapes of the SANS curves depends not only on the concentrations of the molecules but on any solvent conditions that can influence the strength of long-range electrostatic interactions between the molecules. Where these long-range interactions are strong, interparticle effects can be seen at remarkably low concentrations. When the effects are more subtle, the results may be only a slight decrease in apparent R_g and in the molecular weight calculated at the lowest concentrations.

In order to check for the effects of intermolecular interactions, SANS data must be collected at several concentrations. Ideally, at least two concentrations will be found where the $I(q)$ curves scale linearly. Then, in effect, $S(q) = 1$ independent of the number of particles, and the precision of the calculations of shape from $P(q)$ will be optimal, as will extrapolations of $I(q)$ to $S(q)$. However, if $B(q)$ and/or $S(q)$ vary with concentration and cannot be accounted for, the structural properties calculated from $P(q)$ using equation (5) will appear to depend on concentration even if they do not. Furthermore, if the intermolecular interactions are as strong as those shown here for PEG 400 and cytochrome c, then an extrapolation to zero concentration may provide a better estimate of molecular weight.

4.4. Comparisons of data to model structures

4.4.1. Calculating *I*(*q*) from high-resolution structures. SANS is a relatively low-resolution technique; it cannot obtain the resolution equivalent to an X-ray crystal structure. However, when an X-ray crystal or NMR structure of a biological macromolecule is available, it is possible to calculate a model SANS $I(q)$ and an R_g that allow a direct comparison with the SANS data. One widely used program, *CRYSON*, calculates model SANS intensities using spherical harmonics (Svergun *et al.*, 1998). Another treats each type of amino acid residue as a sphere of equivalent scattering length density and size and places the centers of mass of the spheres at the coordinates of the α-carbon (Krueger *et al.*, 1998). This allows for a space filling model of the structure that takes the scattering length density contribution from hydrogen into account. Both types of calculations produce nearly congruent model SANS curves from the same high-resolution structure, and both require specifying the contrast between the protein and solvent in order to match the conditions of the SANS experiment. Here, we restrict our discussion to proteins in simple aqueous solvents consisting of salts at about 100 m*M* in water.

If the high-resolution and solution structures are in fact the same, then the calculated SANS curve normally matches the data very well, and R_g from the model also agrees with the R_g value obtained from the SANS data. However, often a mismatch of the curves occurs at higher q values, *i.e.* those near the background level. In such cases, one may bring the experimental and calculated results to closer agreement in two ways. The first is by adjusting selected parameters consistent with the presence of a surface hydration layer, as can be done with *CRYSON*. However, in practice, similar improvement of the agreement between the model SANS curves and the data can often be obtained in the second manner: by a surprisingly small change in the background subtraction. The resulting difference in the absolute baselines is within the errors that we have described in the preceding sections.

Fig. 6 shows three model SANS curves, along with the measured SANS data from lysozyme at 5 mg ml^{-1} concentration in D$_2$O buffer. The calculation is based on the X-ray structure 6lyz (Protein Data Bank). One of the curves represents the best fit obtained using *CRYSON*, and the parameters from that fit are listed in Table 4, along with parameters from similar fits to the 10 and 20 mg ml^{-1} data. The other two model SANS curves plotted in Fig. 6 were calculated using *XTAL2SAS* (Krueger *et al.*, 1998), but assuming no hydration layer; one curve has no adjustment, and the second, lower one, has been corrected by subtracting 0.0035 cm^{-1}. The latter *XTAL2SAS* curve overlies the *CRYSON* fitted curve.

The *CRYSON* fit assumes a hydration layer of thickness 3 Å. Then, it requires three parameters beyond the X-ray structure to fit the curve: the volume of solvent displaced by the protein, the density of the 3 Å water layer and a background correction. A nearly congruent model curve can be found with a single parameter, adjusting the background with a change well within the standard deviations of the data.

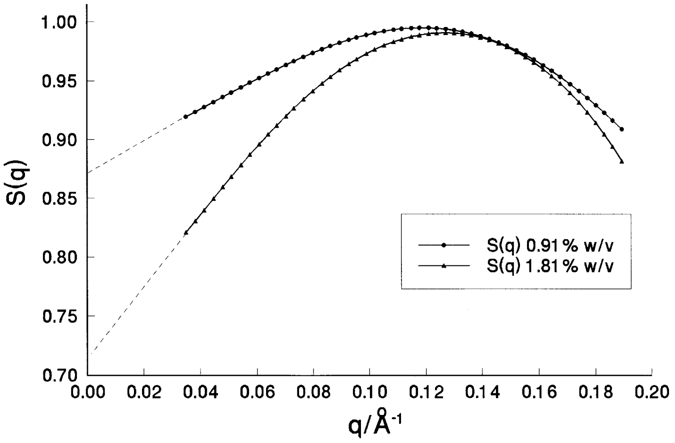


Figure 5
The values of $S(q)$ for cytochrome c from the data of Wu & Chen (1987) obtained using equation (8) and best fitting curves of the data. The protein was assumed to be non-interacting at 0.45%(*w*/*v*), with $S(q) \equiv 1$ at all q.

Table 4
Parameters for *CRYSON* best fit for lysozyme at three concentrations in D_2O.

	5 mg ml^{-1}	10 mg ml^{-1}	20 mg ml^{-1}
Experimental data			
Gunier R_g (Å)	13.0 (1)	12.8 (1)	12.2 (1)
CRYSON best fits assuming 3 Å bound D_2O layer†			
R_g (Å) including water	13.1	13.0	12.7
Mass density‡ (g cm^{-3}) of bound D_2O	1.100	1.107	1.115
Effective volume§ (Å^3)	18 623	18 623	18 539
Background (cm^{-1})	5×10^{-5}	0.0001	0.0017

† Based on Protein Data Bank structure 6lyz. ‡ From $\delta\rho$, the change in scattering length density. D_2O density: liquid at 277 K = 1.105; ice at melting = 1.017. § From R_a, the average displaced solvent volume per atomic group.

Lysozyme is a small protein, and any contrasting surface hydration layer would be quite large relative to the 'core'. Larger proteins would be less effective at testing the presence of a bound water layer with properties differing from the bulk. Within that context, we can examine the *CRYSON* best fit parameters in Table 4, where it is evident that the density of the bound water molecules found from the *CRYSON* fits shows no difference from the density of pure D_2O for all three lysozyme solutions. This result agrees completely with that expected from the descriptive chemistry presented in §4.2.1. However, clearly the model has set an artificial boundary between the two volumes with equal scattering length densities.

The example presented here clearly shows how carefully comparisons must be made between measured SANS curves and model SANS curves calculated from high-resolution X-ray crystal or NMR structures. Furthermore, the conclusions drawn may be software specific.

4.4.2. Low-resolution model structures. When a high-resolution structure is not available, a low-resolution structure can be built from simple shapes of uniform scattering length density, such as a sphere, cylinder or ellipse of rotation. In these cases, the model SANS curves obtained from these structures can be fitted to the data by adjusting the parameters that define the chosen shape and contrast. The results obtained from such models qualitatively match those from the high-resolution structures. In essence, the models provide entirely adequate fits to the $P(q)$ data from scattering by a homogeneous structure bounded by a sharp contrast interface. When additional parameters are used to model inhomogeneous contrast layers, one or another parameter compensates in the fitting so that the final scattering curve is equivalent to a volume with an abrupt change in contrast between the scatterers and the solution. For example, for a sphere modeled by a core region surrounded by a shell that has a scattering length density with any value between that of the solvent and the core, the scaling term will compensate so that the fits are identical to that for a homogeneous sphere alone. The variation of the radius and scaling is equivalent to that of the shell and shows that the scattering appears not to be sensitive to the details of the transition between the different bulk scattering length densities. In addition, the same uncertainties occur for the low-resolution, approximate structures as for the high-resolution structures. Frequently, the best-fit shapes and sizes of the models are sensitive to the position of the baseline.

5. Conclusions

We have outlined and given examples of a number of details that must be considered to obtain valid structural information from SANS data from biomolecules in solution. These details are important because when SANS data from biomolecules in solution are compared with model SANS curves calculated from high-resolution structures from X-ray crystallography they may differ. A simple overall comparison is through the R_g values for each. A difference in R_g can arise for a number of reasons:

(*a*) The solution and crystal structures are not the same.

(*b*) Aggregates in the solution produce a larger average experimental R_g.

(*c*) Interparticle interference causes an apparent shrinkage in R_g.

(*d*) Uncertain H–D exchange relates to uncertainties in the contrast, which has an effect on the calculated R_g for model structures that are being compared with measured SANS data.

(*e*) Bound water with a mass density different from that of bulk water contributes to the scattering and changes R_g.

We have shown that more measurements than have commonly been made are required to correct for incoherent proton scattering, uncertain H–D exchange and intermolecular interactions. However, once these corrections are made, structural parameters derived from the SANS curves

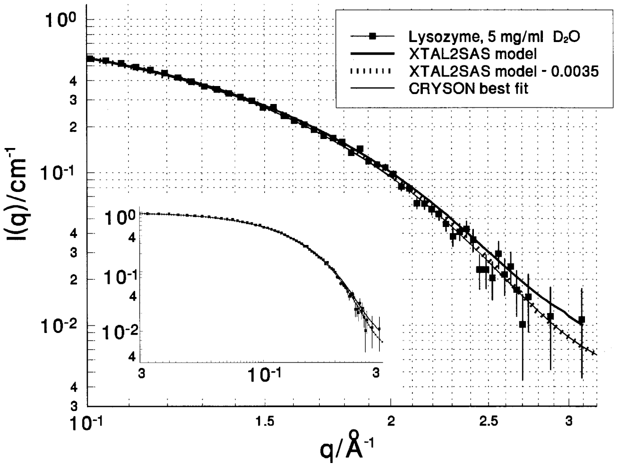


Figure 6
SANS data from 5 mg ml^{-1} lysozyme in D_2O buffer and the resulting model SANS curves from *CRYSON* and *XTAL2SAS* using the structure from the PDB-listed parameters of 6lyz. The *CRYSON* curve represents the best fit to the SANS data assuming a 3 Å bound D_2O layer (fit parameters shown in Table 4), whereas the *XTAL2SAS* model curve assumes no hydration layer. The *XTAL2SAS* curve with a baseline having a constant 0.0035 cm^{-1} subtracted is also shown for comparison. Error bars in the data for $q < 0.2$ Å^{-1} are smaller than the data points. The inset shows the SANS curve in the full measured q range.

calculated from high- or low-resolution structure models can be compared with the SANS data with confidence.

We wish to thank John Barker for a number of insightful discussions. Small-angle neutron scattering experiments were performed on the NG3 and NG7 30 m SANS instruments at the NIST Center for Neutron Research. This equipment was partially supported by the National Science Foundation under agreement No. DMR-9986442.

References

Ball, P. (2003). *Nature* (*London*), **423**, 25–26.

Carsughi, F., May, R. P., Plenteda, R. & Saroun, J. (2000). *J. Appl. Cryst.* **33**, 112–117.

Chalikian, T. V., Totrov, M., Abagyan, R. & Breslauer, K. J. (1996). *J. Mol. Biol.* **260**, 588–603.

Edison, A. S., Weinhold, F. & Markley, J. L. (1995). *J. Am. Chem. Soc.* **117**, 9619–9624.

Efimova, Y. M., van Well, A. A., Hanefeld, U., Wierczinski, B. & Bouwman, W. G. (2005). *J. Radioanal. Nucl. Chem.* **264**, 271–275.

Ghosh, R. E. & Rennie, A. R. (1999). *J. Appl. Cryst.* **32**, 1157–1163.

Glinka, C. J., Barker, J. G., Hammouda, B., Krueger, S., Moyer, J. J. & Orts, W. J. (1998). *J. Appl. Cryst.* **31**, 430–445.

Guinier, A. (1939). *Ann. Phys.* **12**, 161–237.

Hansen, S. (1990). *J. Appl. Cryst.* **23**, 344–346.

Harrison, R. W., Wlodawer, A. & Sjölin, L. (1988). *Acta Cryst.* A**44**, 309–320.

Hayter, J. B. & Penfold, J. (1983). *Colloid Polym. Sci.* **261**, 1022–1030.

Heidorn, D. B. & Trewhella, J. (1988). *Biochemistry*, **27**, 909–915.

Jacrot, B. & Zaccai, G. (1981). *Biopolymers*, **20**, 2413–2426.

Jensen, T. R., Jensen, M. O. S., Reitzel, N., Balashev, K., Peters, G. H., Kjaer, K. & Bjørnholm, T. (2003). *Phys. Rev. Lett.* **90**, 086101.

Kline, S. R. (2006). *J. Appl. Cryst.* **39**, 895–900.

Krueger, S., Groshkova, I., Brown, J., Hoskins, J., McKenney, K. H. & Schwarz, F. P. (1998). *J. Biol. Chem.* **273**, 20001–20008.

Levitt, M. & Park, B. H. (1993). *Structure*, **1**, 223–226.

May, R. P., Ibel, K. & Haas, J. (1982). *J. Appl. Chem.* **15**, 15–19.

Perkins, S. J. (1986). *Eur. J. Biochem.* **157**, 169–180.

Porod, G. (1982). *Small Angle X-ray Scattering*, edited by O. Glatter & O. Kratky, ch. 2. London: Academic Press.

Schowen, K. B. & Schowen, R. L. (1982). *Methods Enzymol.* **87**, 551–606.

Schwendel, D., Hayashi, T., Dahint, R., Pertsin, A., Grunze, M., Steitz, R. & Schreiber, F. (2003). *Langmuir*, **19**, 2284–2293.

Shibayama, M., Nagao, M., Okabe, S. & Karino, T. (2005). *J. Phys. Soc. Jpn*, **74**, 2728–2736.

Squires, G. L. (1996). *Introduction to the Theory of Thermal Neutron Scattering*. Mineola, NY: Dover.

Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.

Svergun, D. I., Richard, S., Koch, M. H. J., Sayers, Z., Kuprin, S. & Zaccai, G. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 2267–2272.

Wu, C.-F. & Chen, S.-H. (1987). *J. Chem. Phys.* **87**, 6199–6205.