Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution

David A. Jacques,[a] J. Mitchell Guss,[a]* Dmitri I. Svergun[b] and Jill Trewhella[a]

[a]School of Molecular Bioscience, The University of Sydney, NSW 2006, Australia, and [b]European Molecular Biology Laboratory, Hamburg Outstation, EMBL c/o DESY, Notkestrasse 85, 22603 Hamburg, Germany

Correspondence e-mail: mitchell.guss@sydney.edu.au

Received 29 December 2011
Accepted 20 March 2012

Small-angle scattering is becoming a mainstream technique for structural molecular biology. As such, it is important to establish guidelines for publication that will ensure that there is adequate reporting of the data and its treatment so that reviewers and readers can independently assess the quality of the data and the basis for any interpretations presented. This article presents a set of preliminary guidelines that emerged after consultation with the IUCr Commission on Small-Angle Scattering and other experts in the field and discusses the rationale for their application. At the 2011 Congress of the IUCr in Madrid, the Commission on Journals agreed to adopt these preliminary guidelines for the presentation of biomolecular structures from small-angle scattering data in IUCr publications. Here, these guidelines are outlined and the reasons for standardizing the way in which small-angle scattering data are presented.

1. Introduction

The last two decades have seen a rapid increase in the use of small-angle scattering for the study of biomolecular structures (Jacques & Trewhella, 2010; Mertens & Svergun, 2010). The explosion in the use of this technique has largely been driven by the increasing desire to characterize biomolecular structures in solution and the availability of easy-to-use software for the analysis and interpretation of small-angle scattering (SAS) data. The latter now also include modelling algorithms for generating three-dimensional models from solution scattering data that provide results in the form of bead or atomic coordinates. To date, no community-agreed set of publication requirements has been available, leading to inconsistencies in which data are reported in publications and to what level of detail. In order to evaluate the interpretation of SAS data, information concerning sample quality, data acquisition and experimental validation are essential, especially when detailed three-dimensional structures are presented. The omission of these important data can lead to inaccurate structural parameters and the generation of erroneous and misleading structural models, the validity of which cannot be independently assessed.

With SAS emerging as a mainstream structural biology technique, and a growing market in commercial instrumentation as well as new SAS beamlines at synchrotron and neutron sources, there has been considerable community drive for the establishment of publication requirements and standards for structural biology applications. The increasing use of SAS in high-throughput efforts (Round *et al.*, 2008; Hura *et al.*, 2009; Grant *et al.*, 2011) also underscores the need for such guidelines. The IUCr, through its Commissions on Small-Angle Scattering and on Journals, has acted to introduce a series of guidelines for the presentation of SAS data in IUCr journals. These guidelines may be found at http://journals.iucr.org/services/sas/. In parallel, a Small-Angle Scattering Task Force has been established to advise the Protein Data Bank on whether models based on SAS data analysis should be deposited and, if so, in what format and with what kinds of supporting data and validation.

Importantly, the guidelines presented here are not being developed to define a quality requirement for SAS experiments that would be acceptable for publication. Rather, the purpose is to establish the way in which SAS experiments should be presented in order to enable a

reviewer and a reader to independently assess the validity of the interpretations made by the authors.

In the present paper, we make the IUCr agreed guidelines broadly available to facilitate their consideration by the research community and potential refinement as appropriate.

2. Sample quality

One of the most celebrated aspects of SAS is that it may be performed on samples without the need for crystals or isotopic labelling (except in the case of neutron contrast variation, where perdeuteration may provide additional information). What is less well appreciated is that small-angle solution scattering data may be acquired and processed from any sample, regardless of the sample quality; as a consequence, without critical evaluation and specific checks the results can be misleading.

The interpretation of solution scattering data in terms of a three-dimensional structural model requires that the solution contains identical monodisperse structures and that the conditions approximate those of infinite dilution. In other words, there is no nonspecific aggregation and no distance correlations between particles such as may occur owing to charge repulsion. Solution scattering data may nonetheless be usefully interpreted in cases where there are associations, mixtures or flexibility and molecular crowding (Rambo & Tainer, 2011; Johansen *et al.*, 2011). In these cases, however, the interpretation will be distinct from interpretation of structural parameters and modelling to represent an individual molecular structure.

In general, SAS patterns reflect not only the structure of individual particles, but also interparticle interactions, with the latter affecting the lowest angle scattering data (Chen & Bendedouch, 1986). SAS is very sensitive to attractive interactions leading to aggregation, showing a rise in the intensity owing to the dependence of the scattering signal on the square of the molecular volume of the scattering particle. Repulsive interactions (*e.g.* between highly charged molecules) tend to diminish the scattering at low angles. When the repulsive or attractive effects are large they are generally easy to spot, as the conditions for a linear Guinier region in the scattering data break down (Guinier, 1938). It is when these effects are at a level such that one still obtains a linear Guinier region but with artificially enhanced or suppressed low-angle scattering data that problems arise. In such a case, the derived parameters and molecular shapes will be too large or too small and more careful analysis is required to avoid being misled. Here, we consider requirements for sample purity and characterization.

2.1. Macromolecular sample purity

As with all biological macromolecular experiments, the purification protocol and an estimate of the final sample purity must be reported. Contamination with high-molecular-weight species, in particular, will bias the data and result in structural parameters and models that are systematically too large. If one in ten molecules (or particles) in the solution have ten times the molecular mass of the molecule of interest, they will account for half of the measured scattering signal and will dominate at the lowest scattering angles. One in ten molecules with five times the molecular mass will contribute 2.5% of the signal. In other words, the degree of contamination that can be tolerated depends on the molecular weight of the contaminating species. Samples that are >99% pure as determined by methods such as SDS–PAGE or the ratio of UV absorbance at 260:280 nm, as appropriate, would generally be adequate. However, it should be appreciated that these methods are qualitative, with SDS–PAGE being insensitive to aggregation (as aggregates are usually dissociated during denaturation by SDS) and UV absorbance at 260:280 nm being most sensitive to nucleotide or nucleic acid contamination. Nevertheless, authors should always provide evidence of the degree of purity of their samples.

2.2. Preparation of solvent blank

Proper subtraction of the scattering arising from solvent is essential to obtain an accurate scattering profile for the macromolecular solute. This is true not only for structural analysis, but also for Kratky (Glatter & Kratky, 1982) analysis, which can provide information on whether a protein is folded and globular, potentially unfolded and flexible, or has flexible regions. Accurate solvent background measurement can be nontrivial, especially for small-angle neutron scattering, where the incoherent scattering from hydrogen in the solvent is large and gives rise to a strong background signal that can be much larger than the macromolecule signal. Dialysis or buffer exchange by size-exclusion chromatography (SEC) are probably the best methods for obtaining a sample of solvent that is 'matched' to the protein and solvent sample. Taking the filtrate from a centrifugal concentration device often yields unmatched solvent blanks owing to the presence of preservative compounds in the membrane (such as glycerol). Solvent mismatch manifests in the high-q data, resulting in either an artificially high or a negative intensity after buffer subtraction. Negative intensity is a physical impossibility, but high intensity at high q can be indicative of sample flexibility, and thus confidence in the solvent subtraction is critical to correct interpretation of scattering data.

2.3. Sample characteristics reported

The nature of the sample (including the molecular mass of the macromolecule of interest with its amino-acid content, including any modifications resulting from its production, which could simply be in the form of a complete sequence and the number and nature of any bound cofactors) and the precise solvent composition, including all additives, must be reported. Additionally, if neutron contrast variation is being undertaken, the level of deuteration achieved and the method by which this value is determined (usually mass spectrometry) must also be reported. All this information allows calculation of the contrast ($\Delta\rho = \rho_{\rm protein} - \rho_{\rm solvent}$, where ρ is the scattering density; Whitten *et al.*, 2008), which is important for experimental validation (see below). Also important for subsequent experimental validation is the concentration of the macromolecule. Usually, protein or nucleic acid concentration is determined by UV spectrophotometry, but in some cases this may be nontrivial to measure, such as when a protein sample is devoid of tryptophan residues or the buffer contains a compound that also absorbs in the UV, such as DTT. Refractometry provides an alternative method to determine the concentration. Refractometry is advantageous in that the refractive index of a protein or nucleic acid is neither dependent on the folded state nor the sequence of the macromolecule. In any event, the macromolecule concentrations in the samples used to collect the scattering data must be explicitly stated, along with the method by which these values are determined.

2.4. Scattering-data-independent measures of sample quality

One of the most important measures of sample quality concerns the evaluation of potential aggregation. While careful treatment of scattering data can yield information regarding aggregation, or possibly the oligomeric state of the sample, an independent measure

of molecular weight provides confidence in the starting sample quality prior to scattering measurement. Dynamic light scattering (DLS) operates over similar concentration and temperature ranges to SAS, but is more sensitive to aggregation. As a dynamic method, DLS is also sensitive to changes in sample viscosity, so high-concentration samples or samples in D_2O may return artificially high molecular weights unless the data have been corrected for viscosity. Multi-angle laser light scattering (MALLS) is also very useful in this context. Because MALLS measurements are usually made immediately following size-exclusion chromatography (SEC), the molecular-weight profile across the elution peak is a powerful method for determining sample molecular weight and polydispersity. If the instrument is connected to a DLS detector (also known as quasi-elastic light scattering or QELS), the measurement will also give an assessment of conformational polydispersity. Samples that dissociate upon dilution are often identified using the SEC-MALLS method by a drop in molecular weight across the elution peak. SAS data collected from such samples need to be treated carefully to demonstrate that no modelling artifacts arising from dissociation or oligomerization result. Often, it is not possible to conduct SEC-MALLS experiments over the same concentration range as SAS experiments. It is possible, therefore, that dissociation effects may be more severe under the conditions of the lower concentration experiment, which is typically the SEC-MALLS experiment as the sample is diluted on the column (Jacques *et al.*, 2009). Such observations can provide clues as to dissociation constants and potentially the biological relevance of macromolecular associations. Owing to the complementary nature of the information provided by DLS and SEC-MALLS, these experiments can greatly improve the confidence in bead or atomistic models derived from SAS data and therefore the data should be presented where available. This argument has been shown to be particularly true for RNA structures (Rambo & Tainer, 2010). The presentation of the SEC-MALLS profile should provide the light scattering from the void volume to the end of the SEC run, thereby informing the reader of the possible scattering contaminants (in particular aggregates) that may be present in the SAS sample.

3. Data acquisition and reduction

As with any reported experiment, details of how the SAS measurement was performed are essential. Of particular importance are the instrument type and configuration. SAS data are acquired from either conventional laboratory-based instruments or dedicated synchrotron beamlines for X-ray scattering and reactor-based or spallation source instruments for neutron scattering. Instrumentation configuration issues that may affect data interpretation include the sample environment (temperature and sample-cell properties, including window material and path length), the wavelength of the incident radiation, the measured q range and the number of detector positions required to obtain this range (especially important for reactor-based SANS experiments) and information required to account for data-smearing effects such as the incident-beam geometry and wavelength spread. In the case of a line-source instrument the beam profile should be provided (either in terms of dimensions of a defined shape, *e.g.* parameters of a trapezoidal profile, or as an intensity plot as a function of q). Smearing effects can be insignificant for X-ray instruments approximating point geometry. In the case of neutron instruments the smearing effects will generally be significant and the beam-aperture dimensions and wavelength spread (cited as a $\Delta\lambda/\lambda$) should be reported along with sample-to-detector distances.

The data-collection strategy, particularly sample-exposure times, must be reported. It is important to monitor radiation damage (particularly at synchrotron sources) and the method by which this damage (or indeed any time-dependent sample deterioration) is monitored must be reported. Typically, radiation damage is detected by the comparison of successive exposures, with sample deterioration often manifesting as a change in scattering intensity as a function of time (generally an increase in scattering intensity at low q as covalent bonds are broken by free radicals, resulting in unfolding and non-specific aggregation). Radiation damage can be reduced by the addition of radical scavengers (such as DTT, TCEP or ascorbate) or by the use of a flow-cell, which continuously passes the sample through the beam for the duration of the exposure. If any measures are taken to reduce the radiation damage, they should be reported.

Data-reduction protocols and software should also be reported, including the application of corrections for sample absorbance or transmission, detector sensitivity and nonlinearity, data normalization for solvent-scattering subtraction and the method for placing the data on an absolute scale (see below). Importantly, the way that smeared data are treated must be described. Some software packages attempt to desmear data based on a supplied beam profile, while others apply a smearing correction to calculated models in order to fit the data. Inappropriate treatment of smeared data can lead to grossly incorrect models and authors need to demonstrate that the data have been processed correctly.

4. Presentation of scattering data and validation

Once data have been acquired and reduced, data quality must be demonstrated. In crystallography, metrics such as $R_{\rm merge}$, $\langle I/\sigma(I)\rangle$ and data completeness are used to report on data quality. However, in contrast to crystallography, which generally yields diffraction from good-quality samples, SAS data can be acquired from samples of any quality and therefore the data require rigorous evaluation in order to demonstrate that they are interpretable in terms of accurate structural parameters and models. It may be argued that making scattering data publicly available is necessary, or at least desirable. For each specimen where a three-dimensional model is presented, submission of the relevant solvent-subtracted data in ASCII three-column format [q, $I(q)$ and the associated errors] as supplementary materials is suggested.

4.1. Presenting $I(q)$ *versus* q as the primary data

$I(q)$ *versus* q plots, as the unadulterated reduced data, must be reported without artificial truncation of low-q data that can mask the presence of aggregation or interparticle interference effects. $I(q)$ plots should be presented as either linear X–log Y (Fig. 1*a*) or log X–log Y (Fig. 1*c*). The former facilitates the reader evaluating the behaviour of the high-q data, while the latter provides the optimal view for evaluating sample polydispersity. A linear X–linear Y presentation (Fig. 1*d*) does not allow the evaluation of key features in the scattering and is therefore discouraged.

Guinier plots {$\ln[I(q)]$ *versus* q^2; Fig. 1*e*} should also be routinely supplied as these are the most effective at revealing the upturns in intensity at low q that are indicative of aggregation (the smiling Guinier) or downturns that are indicative of interparticle interference (the frowning Guinier). The Guinier linear fit must be shown for a range not exceeding $qR_{\rm g} = 1.3$ for globular scattering particles, and for more asymmetric particles this limit approaches values around 1.0 and as small as 0.8 (Hjelm, 1985). The Guinier plot yields approximations for $R_{\rm g}$ and $I(0)$ from its slope and Y intercept, respectively. It could be useful to report a quantitative estimate of the quality of the Guinier plot, *e.g.* as provided by the *AutoRg* program from the

ATSAS package (Petoukhov *et al.*, 2007). While R_g and $I(0)$ can be calculated more precisely from $P(r)$ analysis (as this method uses all of the data), consistency between the Guinier-derived and $P(r)$-derived values can give confidence in the internal consistency of the scattering profile and it is therefore useful to report both values (Table 1). Additional representations, such as a Kratky plot [$q^2I(q)$ *versus* q; Fig. 1*f*], may also be desirable in order to demonstrate whether the macromolecule is folded and globular or whether it has significant flexibility. The data presented in Fig. 1 were purposefully chosen for their small imperfections (notably at high q), but importantly presented so that the reader can assess potential caveats to any interpretation and a reviewer might ask for revisions.

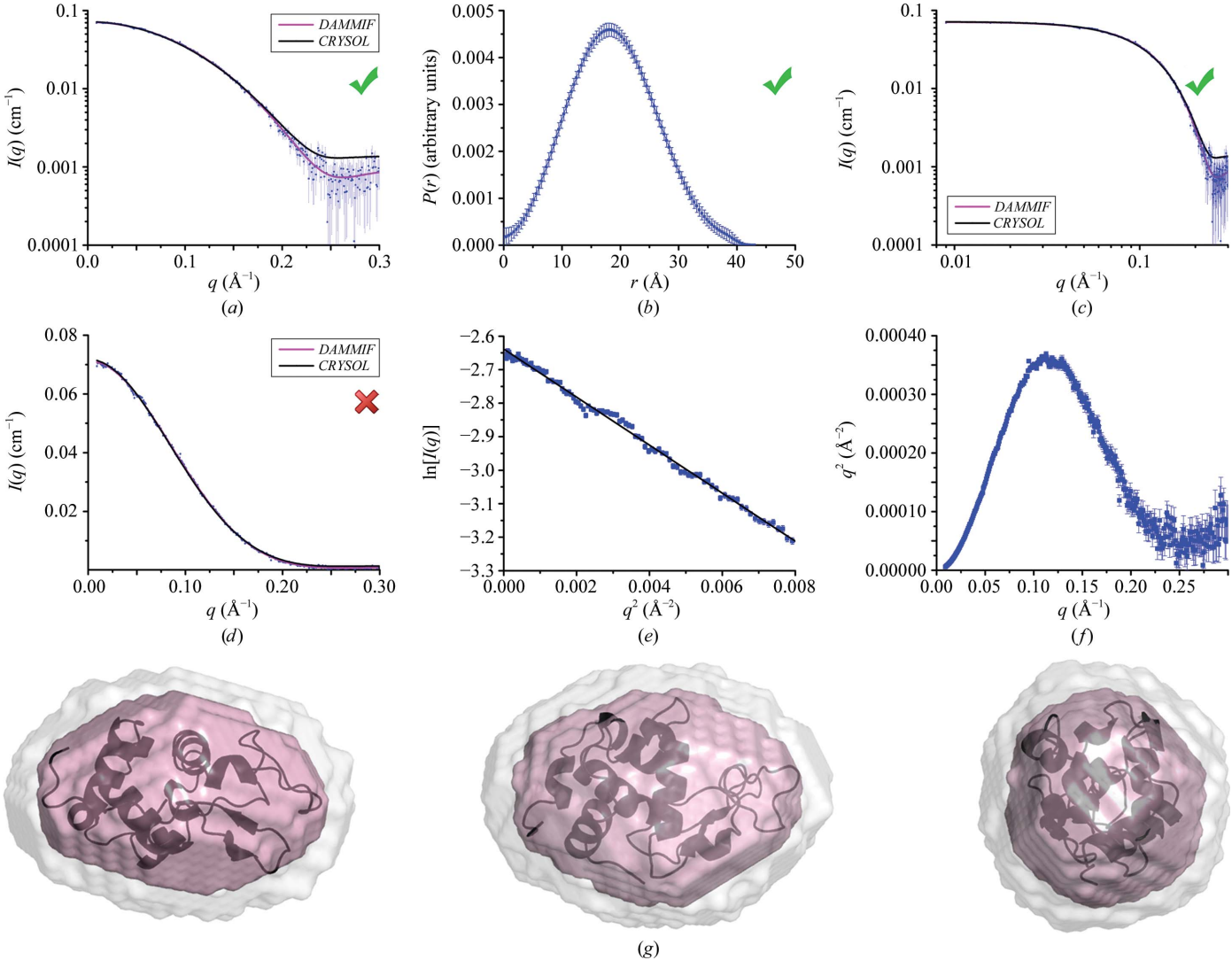

Figure 1
Data were collected on a slit-geometry instrument. Subsequently, all presentations are for smeared data and fits. Scattering data are typically presented as linear *X*–log *Y* plots (*a*) alongside the corresponding $P(r)$ curve (*b*). A log *X*–log *Y* plot (*c*) is also acceptable as it emphasizes the low-angle data that carry the strongest signal and provide the most information regarding the overall shape of the molecule. A sample free from aggregate or interparticle interference will also be flat at small angles, again providing the reader with a rapid diagnostic of data quality. The linear *X*–linear *Y* plot (*d*), however, will obscure both the low-angle information as well as any fits made and must be avoided. Additional representations of the data include the Guinier plot (*e*) and the Kratky plot (*f*). The former provides a rapid diagnostic of sample quality, as deviations from linearity would be indications of either nonspecific aggregation (upturn) or interparticle interference (downturn). The latter provides information as to the folded state of the macromolecule: a fully folded protein would have a parabolic peak followed by convergence at a constant value at high q, while a fully disordered protein would show an increase at high q. If the Porod invariant is used to calculate the molecular mass of the solute, it is necessary to show the Kratky plot to demonstrate that the sample is folded and therefore that the calculation is valid. In this real example, the presented data were used for structural modelling of lysozyme and three orthogonal views of the models generated are presented (*g*). 12 *DAMMIF* calculations (Franke & Svergun, 2009) were performed [a typical fit is presented in magenta in (*a*), (*c*) and (*d*); $\chi^2 = 1.27$] and averaged with *DAMAVER* (Volkov & Svergun, 2003) to produce the averaged and filtered shape shown in magenta in (*g*). It is important to cite the mean normalized spatial discrepancy value and its standard deviation (in this case 0.507 ± 0.009) and whether or not any models in the set were rejected (in this case none) to quantify the degree of similarity among the models generated. In this example, the total volume occupied by the spread of all of the models (aligned for maximum overlap) is shown in grey, with the most-populated volume presented in magenta. The crystal structure of lysozyme has been superposed (black cartoon) on the dummy-atom structure with *SUPCOMB* (Kozin & Svergun, 2001) and its fit to the scattering data calculated with *CRYSOL* [black line in (*a*), (*c*) and (*d*); $\chi^2 = 1.56$; Svergun *et al.*, 1995]. Sources for the discrepancy in the fit for the high-q data should be considered in comments on the interpretation of the data. With the data presented as above, it is possible to see that there is a small upturn at high q in the Kratky plot (*e*), which may be indicative of flexibility (unlikely in the case of lysozyme), a difference between the internal structures of the model (*e.g.* high-resolution features not fully accounted for) and the measured data, or a poor solvent subtraction. The $P(r)$ curve would support the poor subtraction possibility, as the curve does not cleanly approach zero at $r = 0$. With these data available, a reviewer may recommend that the experimenter repeat the measurement before publication, depending on the interpretations made in the manuscript.

4.2. Processed profiles

Even though the goal may be to obtain a molecular model from scattering data, it is useful to provide the Fourier transform of $I(q)$ *versus* q in order to obtain a real-space representation of the data in the form of the probable distribution of the pairwise distances between scattering centres (atoms) within the scattering particle. These $P(r)$ curves (Fig. 1*b*) provide a simple interpretation of the data that can be understood intuitively (Glatter & Kratky, 1982, chapter 5) and also provide evidence for the quality of the data by the manner in which the profile approaches zero at $r = 0$ and $r = D_{\max}$, the maximum linear dimension of the scattering particle. At both limits the approach should be smooth and concave when viewed from above the r axis. Failure of this test for a structured macromolecule at $r = 0$ indicates that there is a problem with the solvent subtraction and at $r = D_{\max}$ can be indicative of aggregation or alternatively that there is significant flexibility in the ensemble of scattering particles. Owing to the finite nature of the measured q range, indirect Fourier methods are used to calculate $P(r)$ from $I(q)$. As $D_{\max}$ is chosen by the experimenter, the ease with which $D_{\max}$ can be unambiguously determined in this process also provides insights into the quality of the data. If a condition $P(D_{\max}) = 0$ is imposed using the indirect transform, it is important that the $P(r)$ function smoothly approaches zero at $D_{\max}$ without a break in the derivative, as the latter may indicate that the $D_{\max}$ value is underestimated.

4.3. Molecular-mass calculations are an important quality check

Determination of the molecular mass or volume of the scattering species is one of the most important parameters to report, as it gives confidence that the scattering is from the molecule of interest without bias from possible weak attractive forces or interparticle interference.

Estimates of molecular mass, M_r, may be obtained directly from the $I(0)$ value if the data are placed on an absolute scale (Orthaber *et al.*, 2000) using

$$M_r = \frac{I(0)N_A}{c(\Delta\rho\nu)^2}, \qquad (1)$$

where N_A is Avagadro's number, c is the sample concentration and ν is the partial specific volume of the macromolecule (see Table 1). Alternatively, a secondary scattering standard such as lysozyme (Krigbaum & Kügler, 1970) may be used to estimate the relationship between $I(0)$ and molecular mass, providing all samples are normalized according to their macromolecular concentrations. This approach, while often employed, assumes that the unknown sample and the standard share a similar contrast and partial specific volume. For samples that have an unusual buffer composition (such as a high concentration of salt or glycerol) or have scattering-length densities significantly different from the standard (such as when the sample contains bound metal cofactors) this assumption breaks down and these factors need to be taken into account.

For globular particles, an alternative estimate of M_r can be made based on the Porod invariant, which provides the excluded particle volume V_p. Empirical calculations show that a relation between M_r and V_p exists which allows one to assess M_r with reasonable accuracy, and tools are available for online calculations (Fischer *et al.*, 2010) or for automated computations (the *AutoPorod* module in the *ATSAS* package http://www.embl-hamburg.de/biosaxs/automation.html).

An experimentally determined value for the molecular mass of the scattering particle in agreement with the expected value (typically within 10%, although an estimate of the uncertainties should be provided) provides confidence that the sample contains monodisperse particles of the expected composition, and analysis of the data to extract structural parameters can proceed.

Table 1
Data-collection and scattering-derived parameters.

Parameters should be reported either normalized by macromolecule concentration or for each point in a concentration series with the sample-concentration values and with details as to how the scattering data were scaled (either to absolute values or relative to a known standard). Where multiple samples are being described, additional columns should be added to provide an easy comparison. The units indicated apply to both X-ray and neutron scattering. [In the case of X-rays, scattering power and contrast values may also be reported as number of electrons (e) and e Å^{-3}, respectively.]

Data-collection parameters	
Instrument	SAXSess (Anton Paar)
Beam geometry	10 mm slit
Wavelength (Å)	1.5418
q range (Å^{-1})	0.009–0.300
Exposure time (min)	60
Concentration range (mg ml^{-1})	2–10
Temperature (K)	283
Structural parameters†	
$I(0)$ (cm^{-1}) [from $P(r)$]	0.114 ± 0.001
R_g (Å) [from $P(r)$]	14.27 ± 0.03
$I(0)$ (cm^{-1}) (from Guinier)	0.112 ± 0.001
R_g (Å) (from Guinier)	14.5 ± 0.1
$D_{\max}$ (Å)	45 ± 3‡
Porod volume estimate (Å^3)	16500 ± 1000
Dry volume calculated from sequence (Å^3)	17570
Molecular-mass determination†	
Partial specific volume (cm^3 g^{-1})	0.724
Contrast ($\Delta\rho \times 10^{10}$ cm^{-2})	3.047
Molecular mass M_r [from $I(0)$]	14100 ± 200
Calculated monomeric M_r from sequence	14300
Software employed	
Primary data reduction	*SAXSquant* 1D
Data processing	*GIFT*
Ab initio analysis	*DAMMIF*
Validation and averaging	*DAMAVER*
Rigid-body modelling	N/A
Computation of model intensities	*CRYSOL*
Three-dimensional graphics representations	*PyMOL*

† Reported for 10 mg ml^{-1} measurement. ‡ $D_{\max}$ is a model parameter in the $P(r)$ calculation and not all programs calculate an uncertainty associated with $D_{\max}$. As such, it is reasonable to not cite an explicit error in $D_{\max}$, although it may be useful to provide some estimate based on the results of $P(r)$ calculations using a range of $D_{\max}$ values.

4.4. Testing the concentration dependence of the scattering data

It is important to determine $I(0)/c$ and R_g at several concentrations of the biomolecule. An increase in these values with concentration is evidence that the sample is undergoing some form of self-association such as oligomerization or aggregation (attractive interactions). On the other hand, a decrease in these values with concentration is evidence of interparticle interference owing to charge repulsion and it may be necessary to adjust the solvent composition to decrease this effect (typically by increasing the ionic strength or adjusting the pH to reduce the particle repulsion). In cases of moderate interactions, it is often possible to extrapolate the scattering to infinite dilution from multiple concentration measurements, assuming that these effects are linearly dependent on the concentration (at low values) of the macromolecule. Whether the data are extrapolated to infinite dilution or a single measurement is used for analysis, data collected at multiple concentrations need to be reported to demonstrate any concentration dependence, or lack thereof, of the observed macromolecular size.

4.5. Neutron contrast variation

In the case of neutron contrast-variation experiments, additional data and analyses are required. The number and the nature of the contrast points needs to be reported (*i.e.* %D_2O solvent values), with

a plot of $I(0)^{1/2}$ (normalized by concentration and exposure time) *versus* solvent scattering density (or $\%D_2O$) that should be linear (reflected at the X axis). This relationship demonstrates that the chosen contrast points provide a sensible level of signal and that the sample is stably monodisperse over the range of solvent conditions chosen. For example, if the sample aggregates at high $\%D_2O$ there will be a consequent deviation from linearity. Molecular-mass calculations from $I(0)$ should be provided for each measured contrast point. Additionally, a Stuhrmann plot of R_g^2 *versus* $\Delta\rho^{-1}$ is desirable as it can provide a model-independent estimate of the R_g values of the individual components as well as that for the overall particle (Ibel & Stuhrmann, 1975). The Sturhmann analysis also provides an estimate of the separation of the centres of mass of the two components, as well as indicating which component is closer to the centre of the complex (Stuhrmann & Kirste, 1967). It is also desirable to present extracted component scattering functions and their resultant $P(r)$ curves to demonstrate the distribution of interatomic vectors within each of the components and between components (the cross-term; Whitten *et al.*, 2008). Tools for these analyses may be found at http://smb-research.mmb.usyd.edu.au/NCVWeb/. A recent example of this type of treatment of neutron scattering data can be found in the investigation of the complex formed between the histidine kinase KinA and its inhibitor Sda (Whitten *et al.*, 2007).

5. Modelling

The conventional analyses described above can give confidence in proceeding to three-dimensional modelling by optimization against scattering data. If a structural model is being put forward for a particular macromolecular system, justification for the specific modelling protocol employed must be provided. A problem frequently encountered when using SAS for structure determination is that of overparameterization. SAS data have an inherently low information content, which leads to the risk of inadvertently introducing more parameters into the model than can be justified. Again drawing parallels with crystallography, the problem of overparameterization during crystal structure refinement has been largely overcome by the use of restraints and the calculation of an R_{free} value. In SAS there is no 'R_{free} equivalent' and so care must be taken to avoid overparameterization. Where a highly parameterized model is reported, the burden is on the author to demonstrate that a simpler model is inadequate to fully explain the data (Jacques & Trewhella, 2010). The example shown in Fig. 1 compares a simple crystal structure fit with that obtained from a dummy-atom reconstruction, but other examples might include the comparison of single rigid-body structures with ensemble models.

Restraints are an effective method for reducing the number of model parameters, but usually these derive from additional experiments, which need to be reported (*e.g.* domain structures, distances from NMR or FRET, symmetry *etc.*). SAS results are at their most robust when modelled in conjunction with information from independent experiments. As such, SAS is often regarded as a powerful complementary technique to high-resolution methods.

When models are presented [including the generation of $P(r)$ curves] authors must report the software used. In the case of three-dimensional modelling, it is important to have a measure of the quality of the fit to the data for any model being proposed. At this time the most common statistical measure used in the modelling of scattering data is χ^2, and this value must be reported for at least the best model. Because χ^2 describes the global goodness-of-fit of the theoretical model scattering to the measured data, it is possible to obtain a low value when fitting data with large errors. In most SAS experiments only counting statistics are used for the calculation of errors. Values of χ^2 of less than 1.0 may arise when counting statistics overestimate the error or when overfitting has occurred. Usually the latter is unlikely, as a smooth function is almost always chosen to fit the data. Such a function is unlikely to result in overfitting, but experimenters should examine their fits to ensure that there is no 'structure' in the calculated curve that might be evidence of overparameterization. Likewise, χ^2 values of above 1.0 occur when counting statistics fail to fully account for the errors in the measurements (*e.g.* when there is a systematic error that is not accounted for in the error model derived from counting statistics alone). This situation is of greatest concern at synchrotron SAXS beamlines, where excellent counting statistics are more easily obtained. In these situations, data reduction is of critical importance to avoid the introduction of systematic errors into the scattering profile. Of course, the most likely explanation for χ^2 values greater than 1.0 is that the proposed model does not fully explain the data. In the event that a model is being proposed where the χ^2 value is significantly greater or less than 1.0, the author must explain why the structural interpretation is valid.

As the absolute value of χ^2 may be somewhat misleading (particularly when comparing the quality of models obtained from different data), a plot of the model fit to the experimental $I(q)$ *versus* q must be shown for at least the best model. This plot allows a qualitative judgment to be made as to the goodness-of-fit to the data and can highlight specific regions of poor fit to the scattering profile. This information may have important implications regarding the accuracy of the final model (an example of local poor fit is shown in Fig. 1*a*).

One consequence of the rotational averaging in small-angle solution scattering data is the possibility that multiple non-unique solutions may be obtained to any modelling calculation. Authors must endeavour to describe the degree of ambiguity of any shape reconstruction. One way to report this ambiguity is by the normalized spatial discrepancy values obtained through clustering or averaging of individual models (Volkov & Svergun, 2003). Additionally, if modelling calculations generate multiple distinct populations of solutions that fit the data equally well, each of these populations should be described. Alternatively, if only one solution is presented, justification for the rejection of other solutions must be made.

Perhaps the most powerful use of SAS is to combine atomic models for individual domains and to use this information to represent the global structure using rigid-body modelling (Petoukhov & Svergun, 2005). Where authors are reporting rigid-body models, a description of how the starting structures were obtained must be provided (*e.g.* crystallography, NMR or homology models). Additionally, any other modelling assumptions that have been made (such as distance restraints, disorder and symmetry) need to be detailed and justified.

6. Concluding remarks

These guidelines largely focus on those SAS experiments that have been used to produce atomic coordinates, whether they are dummy-atom (bead) models or atomic positions from rigid-body calculations. While such structures can be produced relatively easily and may be visually appealing, it is of the utmost importance that authors and readers appreciate the accuracy and limitations of these models, the appropriateness of the modelling techniques employed and therefore the validity of any conclusions drawn. These guidelines form the foundation of what will hopefully be an evolving process of standardizing the way in which structural biology is reported from small-angle scattering experiments.

References

Chen, S.-H. & Bendedouch, D. (1986). *Methods Enzymol.* **130**, 79–116.
Fischer, H., de Oliveira Neto, M., Napolitano, H. B., Polikarpov, I. & Craievich, A. F. (2010). *J. Appl. Cryst.* **43**, 101–109.
Franke, D. & Svergun, D. I. (2009). *J. Appl. Cryst.* **42**, 342–346.
Glatter, O. & Kratky, O. (1982). *Small-Angle X-ray Scattering*. London, New York: Academic Press.
Grant, T. D., Luft, J. R., Wolfley, J. R., Tsuruta, H., Martel, A., Montelione, G. T. & Snell, E. H. (2011). *Biopolymers*, **95**, 517–530.
Guinier, A. (1938). *C. R. Hebd. Seances Acad. Sci.* **206**, 1374–1376.
Hjelm, R. P. (1985). *J. Appl. Cryst.* **18**, 452–460.
Hura, G. L., Menon, A. L., Hammel, M., Rambo, R. P., Poole, F. L., Tsutakawa, S. E., Jenney, F. E., Classen, S., Frankel, K. A., Hopkins, R. C., Yang, S. J., Scott, J. W., Dillard, B. D., Adams, M. W. & Tainer, J. A. (2009). *Nature Methods*, **6**, 606–612.
Ibel, K. & Stuhrmann, H. B. (1975). *J. Mol. Biol.* **93**, 255–265.
Jacques, D. A., Streamer, M., Rowland, S. L., King, G. F., Guss, J. M., Trewhella, J. & Langley, D. B. (2009). *Acta Cryst.* D**65**, 574–581.
Jacques, D. A. & Trewhella, J. (2010). *Protein Sci.* **19**, 642–657.
Johansen, D., Jeffries, C. M., Hammouda, B., Trewhella, J. & Goldenberg, D. P. (2011). *Biophys. J.* **100**, 1120–1128.
Kozin, M. B. & Svergun, D. I. (2001). *J. Appl. Cryst.* **34**, 33–41.
Krigbaum, W. R. & Kügler, F. R. (1970). *Biochemistry*, **9**, 1216–1223.
Mertens, H. D. & Svergun, D. I. (2010). *J. Struct. Biol.* **172**, 128–141.
Orthaber, D., Bergmann, A. & Glatter, O. (2000). *J. Appl. Cryst.* **33**, 218–225.
Petoukhov, M. V., Konarev, P. V., Kikhney, A. G. & Svergun, D. I. (2007). *J. Appl. Cryst.* **40**, s223–s228.
Petoukhov, M. V. & Svergun, D. I. (2005). *Biophys. J.* **89**, 1237–1250.
Rambo, R. P. & Tainer, J. A. (2010). *RNA*, **16**, 638–646.
Rambo, R. P. & Tainer, J. A. (2011). *Biopolymers*, **95**, 559–571.
Round, A. R., Franke, D., Moritz, S., Huchler, R., Fritsche, M., Malthan, D., Klaering, R., Svergun, D. I. & Roessle, M. (2008). *J. Appl. Cryst.* **41**, 913–917.
Stuhrmann, H. B. & Kirste, R. G. (1967). *Z. Phys. Chem. (Frankfurt am Main)*, **56**, 334–337.
Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.
Volkov, V. V. & Svergun, D. I. (2003). *J. Appl. Cryst.* **36**, 860–864.
Whitten, A. E., Cai, S. & Trewhella, J. (2008). *J. Appl. Cryst.* **41**, 222–226.
Whitten, A. E., Jacques, D. A., Hammouda, B., Hanley, T., King, G. F., Guss, J. M., Trewhella, J. & Langley, D. B. (2007). *J. Mol. Biol.* **368**, 407–420.