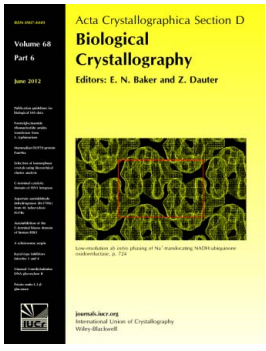
Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449
Editors: **E. N. Baker and Z. Dauter**

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

David A. Jacques, J. Mitchell Guss, Dmitri I. Svergun and Jill Trewhella

Acta Cryst. (2012). D**68**, 620–626



Acta Crystallographica Section D: Biological Crystallography welcomes the submission of papers covering any aspect of structural biology, with a particular emphasis on the structures of biological macromolecules and the methods used to determine them. Reports on new protein structures are particularly encouraged, as are structure–function papers that could include crystallographic binding studies, or structural analysis of mutants or other modified forms of a known protein structure. The key criterion is that such papers should present new insights into biology, chemistry or structure. Papers on crystallographic methods should be oriented towards biological crystallography, and may include new approaches to any aspect of structure determination or analysis. Papers on the crystallization of biological molecules will be accepted providing that these focus on new methods or other features that are of general importance or applicability.

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.

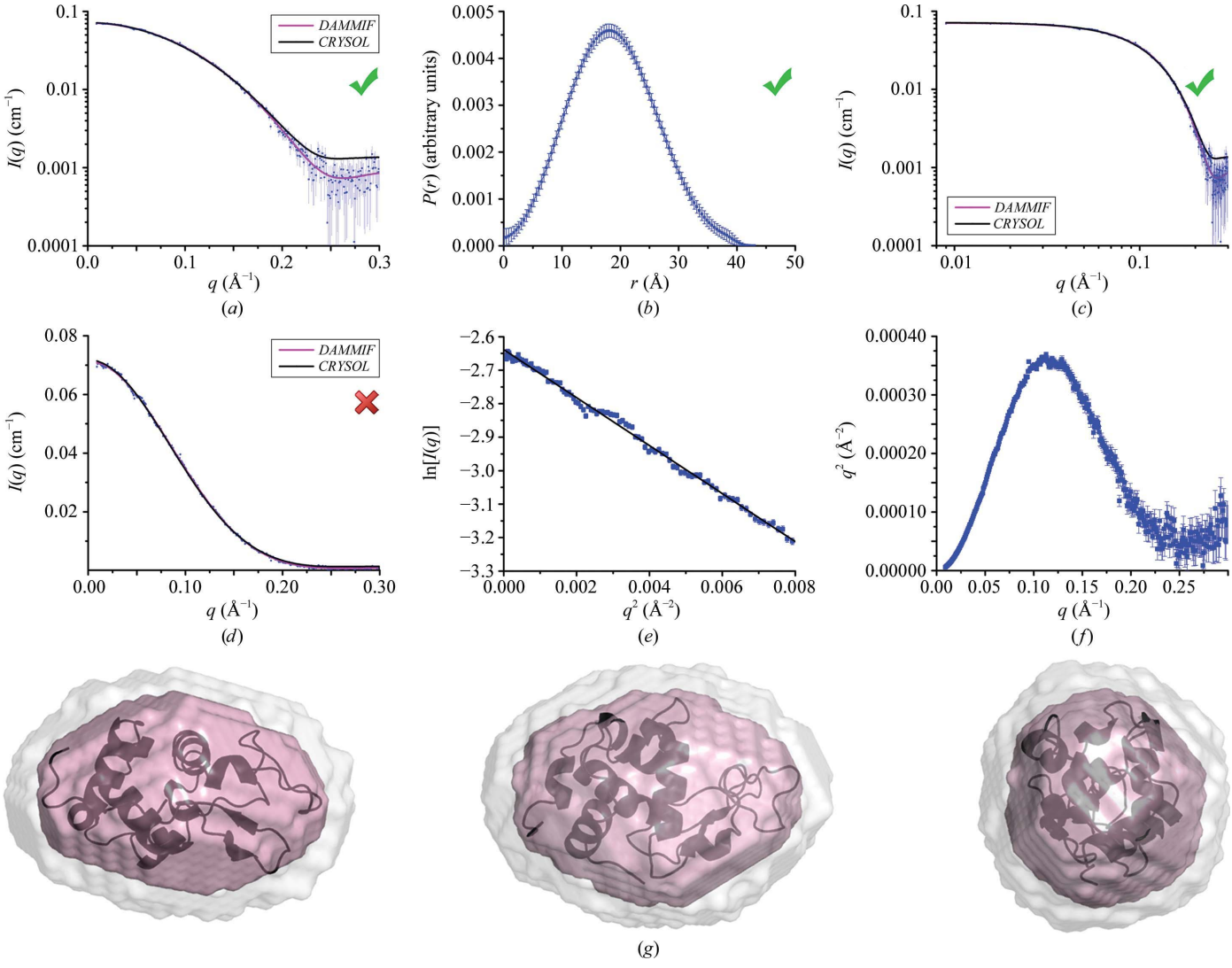


Figure 1
Data were collected on a slit-geometry instrument. Subsequently, all presentations are for smeared data and fits. Scattering data are typically presented as linear X–log Y plots (*a*) alongside the corresponding $P(r)$ curve (*b*). A log X–log Y plot (*c*) is also acceptable as it emphasizes 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

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

where N_A is Avagadro's number, c is the sample concentration and v 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.

Structure
Meeting Review

Report of the wwPDB Small-Angle Scattering Task Force: Data Requirements for Biomolecular Modeling and the PDB

Jill Trewhella,[1,*] Wayne A. Hendrickson,[2] Gerard J. Kleywegt,[3] Andrej Sali,[4] Mamoru Sato,[5] Torsten Schwede,[6,7] Dmitri I. Svergun,[8] John A. Tainer,[9,10] John Westbrook,[11] and Helen M. Berman[11]
[1]School of Molecular Bioscience, The University of Sydney, NSW 2006, Australia
[2]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA
[3]European Molecular Biology Laboratory–European Bioinformatics Institute, Cambridge CB10 1SD, UK
[4]Departments of Bioengineering and Therapeutic Sciences, and Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, CA 94143, USA
[5]Graduate School of Medical Life Science, Yokohama City University, Yokohama, Kanagawa Prefecture 236-0027, Japan
[6]Biozentrum, Universität Basel, University of Basel, 4003 Basel, Switzerland
[7]SIB Swiss Institute of Bioinformatics, 4056 Basel, Switzerland
[8]European Molecular Biology Laboratory, Hamburg Outstation, 22603 Hamburg, Germany
[9]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94704, USA
[10]Department of Integrative Structural and Computational Biology, The Skaggs Institute for Chemical Biology, The Scripps Research Institute, LaJolla, CA 92037, USA
[11]Department of Chemistry and Chemical Biology, Rutgers University, New Brunswick, NJ 07102, USA
*Correspondence: jill.trewhella@sydney.edu.au
http://dx.doi.org/10.1016/j.str.2013.04.020

This report presents the conclusions of the July 12–13, 2012 meeting of the Small-Angle Scattering Task Force of the worldwide Protein Data Bank (wwPDB; Berman et al., 2003) at Rutgers University in New Brunswick, New Jersey. The task force includes experts in small-angle scattering (SAS), crystallography, data archiving, and molecular modeling who met to consider questions regarding the contributions of SAS to modern structural biology. Recognizing there is a rapidly growing community of structural biology researchers acquiring and interpreting SAS data in terms of increasingly sophisticated molecular models, the task force recommends that (1) a global repository is needed that holds standard format X-ray and neutron SAS data that is searchable and freely accessible for download; (2) a standard dictionary is required for definitions of terms for data collection and for managing the SAS data repository; (3) options should be provided for including in the repository SAS-derived shape and atomistic models based on rigid-body refinement against SAS data along with specific information regarding the uniqueness and uncertainty of the model, and the protocol used to obtain it; (4) criteria need to be agreed upon for assessment of the quality of deposited SAS data and the accuracy of SAS-derived models, and the extent to which a given model fits the SAS data; (5) with the increasing diversity of structural biology data and models being generated, archiving options for models derived from diverse data will be required; and (6) thought leaders from the various structural biology disciplines should jointly define what to archive in the PDB and what complementary archives might be needed, taking into account both scientific needs and funding.

Introduction to Small-Angle Scattering: What Can We Learn?

Structural analysis of biologic molecules using small-angle scattering (SAS) is increasingly commonplace, as reflected in the more than tripling of the number of biological SAS publications over the past 10 years (from 105 in 2002 to 355 in 2011). Most publications reporting SAS data contain a three-dimensional (3D) model of some kind, either a shape model or an atomistic representation. The rising interest in SAS has multiple drivers. It enables the determination of precise and accurate structural parameters for biomolecules in solution over a broad size range—tens to thousands of angstroms (Jacques and Trewhella, 2010; Mertens and Svergun, 2010; Rambo and Tainer, 2010). As structural biologists target larger, more complex, and often partly flexible systems, SAS has become a tool of choice to furnish an initial model, albeit limited in resolution, that can provide novel insights into function (Christie et al., 2012; Jacques et al., 2008; Morgan et al., 2011; Nishimura et al., 2009; Rodrigues et al., 2012; Schiering et al., 2011; Whitten et al., 2008; Williams et al., 2009).

With the increased accessibility of small-angle X-ray scattering (SAXS) instruments at synchrotrons, more crystallographers are routinely acquiring SAXS data on the object of their investigations. With the automation currently available, it is becoming practical to acquire SAXS data over a range of conditions, for example in screening crystallization trials to determine conditions under which the target for crystallization is soluble as a mono-disperse species. Furthermore, given the many samples being prepared for both the Protein Structure Initiative and for structural biology in many individual research labs, SAXS efficiently provides solution structural information. For example, in a systematic study of 50 proteins from *Pyrococcus furiosus*,

SAXS analysis was used to determine whether proteins were aggregated or unfolded, to define global structural parameters and oligomeric states for most samples, to identify shapes and similar structures for 25 unknown structures, and to determine molecular envelopes for 41 proteins (Hura et al., 2009).

The growth in the number of small-angle neutron scattering (SANS) experiments lags that for SAXS for a number of reasons: sample sizes are at least an order of magnitude larger and contrast variation requires multiple samples with deuterium labeling; the much lower neutron fluxes achievable compared to X-rays leads to considerably lower signal-to-noise ratios, even with the larger sample sizes; and neutron beams of sufficient intensity for SANS can only be obtained at research reactors or accelerator-based spallation sources that are far fewer in number than synchrotrons. Nonetheless, if the scientific motivation is strong enough for the experiment, SANS with contrast variation provides uniquely valuable information concerning the quaternary structure of biomolecular complexes in solution.

This report concerns issues relating to the archiving of models derived using SAS data, the necessary accompanying data with criteria for assessing data quality, and model validation. In this context it is important to recognize that the assessment of SAS data quality depends to some extent on the specific experiment and questions being asked. The focus here is on experiments aimed at characterizing macromolecular shape and assembly, and/or fitting atomic models to SAS data. Other classes of experiments, such as those aimed a monitoring biophysical processes (e.g., folding, flexibility, filament formation, or overall structural changes), will have overlapping but also distinct criteria.

Structural Information Encoded in the Small-Angle Solution Scattering Pattern and Quantitative Interpretation

The modeling of three-dimensional structures based on SAS profiles is limited by the information content of the SAS pattern, which is essentially one-dimensional and relates to the pairwise distances between scattering centers (atoms) within the macromolecule and their relative scattering powers. Hence, the question of uniqueness always needs to be addressed when assessing 3D models derived from SAS data (*i.e.*, more than one 3D shape may result in the same one-dimensional scattering pattern). The SAS profile (generally expressed as $I(q)$ versus q, where $q = (4\pi\sin\theta)/\lambda$, 2θ is the scattering angle and λ the wavelength of the radiation) can be interpreted in terms of the shape of the scattering object and the distribution of scattering density within that shape. The resolution limit of the solution SAS measurement (typically of the order of 10 Å) is compounded by rotational averaging due to tumbling motions of biomolecules. If there is an ensemble of conformers present or flexibility, the measured profile represents the population-weighted average structure over the measurement period. To interpret SAS data in terms of a single 3D model, it is essential that the solutions be highly pure, monodisperse, and contain identical particles.

In their 1955 monograph, Guinier and Fournet (Guinier and Fournet, 1955) predicted that SAS would be most powerful in its application to the study of biologic macromolecules because, unlike synthetic polymers, they fold into well-defined structures that can meet the stringent requirements of purity and monodispersity necessary for accurate structural interpretation of SAS data. The early developments in quantitative interpretation of SAS data are described by many of the pioneers in Glatter and Kratky's definitive text (Glatter and Kratky, 1982). In the 1930s Guinier showed that the lowest-angle scattering data could provide estimates of the radius of gyration (R_g) and forward scattering intensity ($I(0)$) that gave measures of relative compactness and molecular weight, respectively, of a particle in solution. Other pioneers since have defined additional important and useful relationships, e.g., the Kratky plot ($q^2I(q)$ versus q) for distinguishing folded, unfolded, and flexible molecules, and estimating molecular volumes; Porod's law to describe the asymptote of the scattering intensity $I(q)$ for large q values. In the 1970s, Glatter developed the now standard indirect methods for Fourier transforming experimental $I(q)$ (measured over a finite q-range) to obtain $P(r)$. $P(r)$ is the frequency distribution of distances (r) between scattering centers (atoms) within the scattering molecule weighted by the product of the scattering power at each scattering center and is thus the real space interpretation of I(q). $P(r)$ transformation is often used to generate smoothed scattering profiles for modeling. The zeroth and second moments of $P(r)$ also yield $I(0)$ and R_g values generally with higher precision than Guinier analysis because $P(r)$ is determined using the full measurement range for $I(q)$. All of these early analyses are commonly used in the modern software packages available for SAS data analysis.

While the attention SAS is enjoying today can be attributed both to advances in methodology and changes in the focus of structural biologists, perhaps most influential in the explosion of interest has been the development of easy-to-use SAXS and SANS data interpretation tools, especially the capacity for 3D structural modeling. Figure 1 provides a roadmap for modern SAS data analysis. The much cited and broadly used ATSAS SAS data acquisition and analysis package (Petoukhov et al., 2012) provides the most comprehensive set of SAS data processing and interpretation tools, including those for 3D modeling.

Small-Angle Scattering and 3D Modeling

There are two classes of 3D models that are most frequently generated from SAS data. One class is the "shape" or "ab initio" model where a molecular envelope is generated solely from the SAS data with minimal assumptions (generally continuity and compactness.). These models are commonly represented as arrangements of beads or dummy residues within a defined volume. The second class comprises atomistic models that incorporate high-resolution structural components from X-ray crystallography or NMR spectroscopy and rigid-body refinement against SAS data.

Recent work has demonstrated that significant improvements of NMR-based solution structures can be obtained if the NMR data are co-refined against SAS data (Grishaev et al., 2005, 2008). The success of this approach derives from the long-range distance and translational restraints from the SAS data that complement the short-range distance and orientational restraints derived from the nuclear magnetic resonance (NMR) experiments. Combined use of NMR and SAXS data in model refinement is thus especially powerful for multidomain or multi-subunit structures where NMR restraints are often

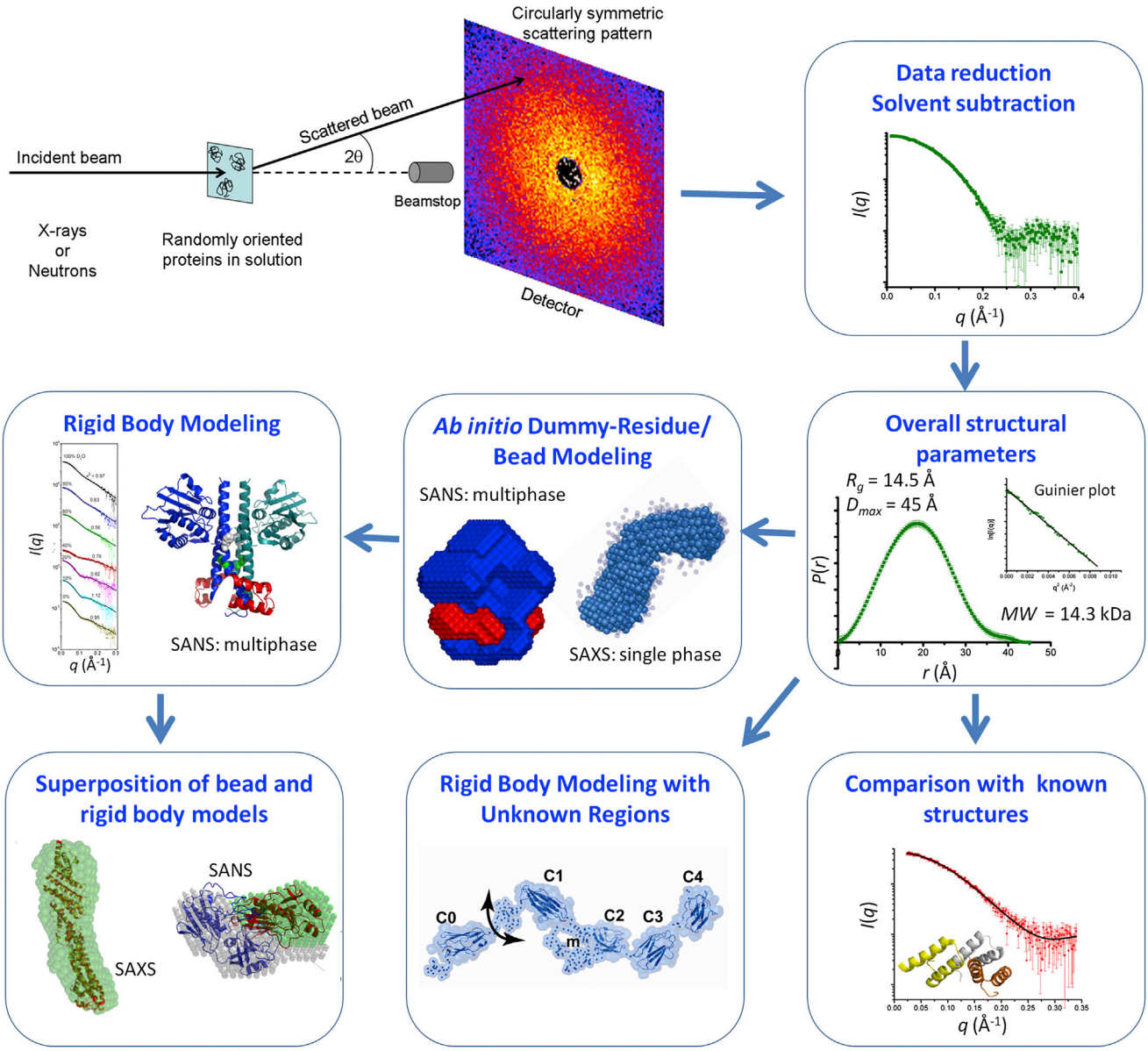


Figure 1. Roadmap of SAS Data Collection and Analysis

Scattering data are measured for a biologic macromolecule in solution on a two-dimensional detector as a circularly symmetric pattern. Data reduction (e.g., corrections for detector sensitivity, linearity, and circular averaging) yields a one-dimensional scattering profile for the macromolecule after subtraction of the solvent contribution to the scattering. The resultant SAS profile can be analyzed to provide overall structural parameters (R_g and molecular weight, *MW*) and *P*(*r*) versus *r* (which also yields the maximum dimension D_{max}). After validation that the scattering particle has the expected MW, comparison can be made with a scattering profile calculated from a PDB coordinate file. Ab initio methods can provide bead or dummy-residue models indicating the shape of the macromolecule. In cases where structures of domains or subunits are known, rigid-body refinement can provide an atomistic model. SAXS data enable single phase modeling, while contrast variation data from SANS experiments enable multiphase modeling. If there are regions of the molecule of unknown structure, these can be modeled using a combination of rigid-body/dummy-residue modeling. Superposition of bead and rigid-body models is one form of model validation.

insufficient to accurately define the domain or subunit interfaces. The improvement in agreement between NMR/SAXS and crystal structures (compared with NMR-only and crystal structures) has been quantified for one of the largest single-polypeptide structures to be solved by solution NMR methods (82 kDa malate synthase G, Protein Data Bank [PDB] accession code 2JQX; Grishaev et al., 2008). Analysis revealed that the improvement was due primarily to the influence of the medium-angle scattering data.

Atomistic models are obtained by rigid-body refinement against a SAXS or SANS data set (Jacques et al., 2008, 2011; Nicastro et al., 2010; Putnam et al., 2007; Whitten et al., 2008). The atomistic information is generally taken from crystallographic, NMR, or homology models representing domains or subunits whose positions and orientations are refined against SAXS or SANS data, or both. Additional constraints may be applied from other experiments such as distance constraints from fluorescence resonance energy transfer (FRET) and

cross-linking studies. The final model will include regions of unknown structure or a structural interface where there is no reliable atomistic information (e.g., linkages between domains, interfaces between domains or subunits).

Aside from NMR/SAXS co-refinement or rigid-body modeling of atomistic models against SAS data, tools for integrative modeling using data from different sources are under development. The integrative modeling platform (IMP) (Russel et al., 2012; Schneidman-Duhovny et al., 2012) has been developed as part of the National Center for Dynamic Interactome Research for integrating various data (atomic, course-grained, SAXS, electron microscopy, proteomics, cross-linking, FRET, etc.) Protein structure prediction with Rosetta (http://boinc.bakerlab.org/; Leaver-Fay et al., 2011) also is increasingly providing for data integration as are data-driven docking approaches such as HADDOCK (Karaca and Bonvin, 2012).

Benefits of Making SAS Data and SAS-Derived Models Publicly Accessible

There are multiple and compelling reasons for making SAS data and SAS-derived models publicly accessible for evaluation and utilization of data and models, as well as the development of improved computational methods for analysis and interpretation. Even in the absence of a proposed model, SAS data provide useful information on the solution state of a system (e.g., oligomerization state and monodispersity). The shape model, as the minimalist 3D structural interpretation of the measured scattering profile, can provide useful insights and is helpful for comparing the solution structure to crystallographic, NMR, and homology models. In the case of models deposited in the PDB for which SAS data have made an essential contribution to the final result, such as in combined NMR-SAS structural refinement, the SAS data need to be made available.

Combinations of methods are increasingly being used to study biomolecular structures, especially as we strive to define more complex assemblies such as molecular machines or even cellular components (Alber et al., 2008). It is these much more complex structures that are at today's structural biology frontier, where we seek to understand biologic function at the molecular level with as much detail as possible. Many of the approaches to studying these more complex systems utilize relatively low-resolution and even low-information content methods. Examples of low-information content methods, when compared to high-resolution crystallography, are those that provide information primarily on shape or proximity of component molecules (e.g., SAS, FRET, double electron-electron resonance [DEER], mass spectrometry, hydrogen exchange, cross-linking, affinity purification, electron tomography, soft X-ray tomography, etc.). Models to interpret such data may include components that are atomistic, e.g., rigid-body crystallographic or NMR structures fitted into electron microscopy maps or optimized initially to SAS data, but overall they will not be uniformly detailed or accurate. Nonetheless, given the investment in developing these kinds of models using such diverse data, it is important that they are archived and available to the broader community for evaluation, testing, and methods development, as well as for designing hypotheses to drive further experiments aimed at advancing our understanding of the system.

Current State of Repositories

The first SAS-derived entries were deposited to the PDB archive in 1999, and a detailed "REMARK 265" was created to report SAS experimental details (Boehm et al., 1999). These structures are atomistic models determined either by rigid-body fitting of existing X-ray or NMR structures or by computational modeling. Currently, no SAS bead models are released in the PDB archive or on policy hold. Acceptance of atomistic models for which the only experimental input is SAS data was interrupted in 2009, and these kinds of SAS-derived structures deposited subsequently have been placed on policy hold pending the recommendations of the wwPDB SAS Task Force. Structures determined using SAS data but substantially derived from other experimental methods continue to be accepted and incorporated into the PDB. The majority of these entries have been determined using SAS and solution NMR, with a few structures determined using SAS and electron microscopy.

An independent web-accessible database for storing and distributing peer-reviewed SAXS data is Bioisis, available at http://www.biosis.net, which may complement and inform future efforts to archive SAS data. Every entry is given a unique identification code that corresponds to a SAS experiment with a sample in a particular solution state. The deposition requires an explicit description of solution conditions (e.g., pH, monovalent and divalent ion concentrations, additives, etc.) and instrument parameters (e.g., wavelength, exposure times, and source). A Bioisis deposition does not necessarily require a 3D model because some experiments may be designed for nonmodeling purposes such as unfolding studies of protein or RNA samples. An entry may be composed of more than one SAXS curve and Bioisis is capable of storing multiple SAXS curves to allow for an assessment of a non-unit structure factor arising from interparticle interference due to long-range distance correlations in the sample. Bioisis was designed with the intent of allowing a depositor to upload the entire set of SAXS curves that led to the final conclusion or model. Often in the literature, analysis is performed using both dummy-residue models and atomistic models. Bioisis allows a single entry to contain multiple models derived from dummy residues or atomistic ensemble models. If a dummy-residue model is deposited, Bioisis requires the unaveraged models as well as the averaged model for deposition. A deposition may be downloaded as a Zip file containing all the experimental information and models. Bioisis restricts deposits to SAXS experiments that have been published in peer-reviewed journals, providing researchers with the SAXS data used to support a given published interpretation.

Recommendations of the Task Force

A Global Data Repository Is Needed that Holds Standard Format X-Ray and Neutron SAS Data that Are Searchable and Freely Accessible for Download

A globally accessible archive or repository for deposition of SAS data in a standard format with sufficient information regarding the sample, the SAS instrument geometry, data acquisition, and reduction would provide researchers with a wealth of information about the solution state of specific systems. For NMR structures that are in the PDB and have been obtained by corefinement with SAXS data, the SAXS data should be made

available and in a standard format; either via a link to a dedicated archive for SAS data or as part of the PDB entry.

A Standard Dictionary Is Required for Definitions of Terms for Data Collection and for Managing the SAS Data Repository

A prerequisite for the envisioned internationally accessible archive for SAS data is an agreed set of definitions for what data would be required for a submission and in what format. The IUCr Small-Angle Scattering and Journals commissions have developed and accepted a set of draft guidelines for the publication of SAS data (Jacques et al., 2012; http://journals.iucr.org/services/sas/). These recommendations, along with later recommendations developed by the canSAS 1D Formats Working Group (http://www.small-angle.ac.uk/small-angle/Formats/canSAS-1D-1-0.html) provide an excellent starting set of requirements. The following requirements are consistent with the IUCr guidelines, with some additional requirements and specifications of the format for the scattering data.

For an international SAS archive, the solvent-subtracted SAS data must be provided, along with all of the measurements used to obtain them. The SAS data in an ASCII three-column format (q, $I(q)$, and associated error $Er(q)$) would be the simplest option. For shape models, an additional column would contain the model $I(q)$ for each q value used in the experiment. All SAS intensity data should be on an absolute scale in units of cm^{-1} with the method for determination of the absolute scaling specified, e.g., by reference to a well-characterized scattering standard, such as H_2O or a known protein, or relative to incident beam flux. In the case of SANS contrast variation experiments, data for each contrast point measured should be deposited.

Ideally, SAS data measured at multiple concentrations would be provided as evidence for the absence of interparticle interference arising from long-range distance correlations or concentration-dependent aggregation that would bias the structural interpretation. If final analysis is carried out on SAS data that has been extrapolated to infinite dilution, or merged in some way from multiple measurements, this processed data set should be provided along with the original measured data and the protocol by which the extrapolated or merged data set was obtained.

In addition to the SAS data, information regarding data acquisition and reduction should be specified, including the wavelength of the radiation and any wavelength dispersion, detector characteristics, basis for error estimates (Poisson counting statistics or not), methods for detector sensitivity and linearity corrections, the geometry of the SAS instrument, and radiation source. Data smearing parameters resulting from the geometry of the instrument and/or a wavelength distribution in the incident radiation must be specified. Where the smearing effects are significant and de-smeared data were used to develop models, the de-smeared data should be provided in the same format as the measured (smeared) data.

Details of the sample are essential, addressing sample content including amino acid or nucleic acid sequences; composition of any ligands, cofactors, or modifications; sample purity; solvent composition and pH; concentration of the biomolecules (and the means by which it was determined); and sample temperature for measurement. In the case of SANS contrast variation experiments, accurate percentage deuteration of each biomolecular component (e.g., from mass spectrometry) and the solvent (e.g., from densitometry, weighing) must be included with information on how they were determined.

Previous work by the SAS community and the IUCr led to a consensus on an ASCII format for one-dimensional SAS data that includes a self-describing header containing relevant information about the sample and instrumental conditions followed by raw or reduced data in a tabular form. This format called sasCIF was implemented as an extension of the core CIF (crystallographic information File) dictionary (Malfois and Svergun, 2000). This dictionary should be reviewed and updated as needed to provide the basis for the SAS data collection dictionary.

Options Should Be Provided for Including in a Repository SAS-Derived Shape—e.g., Bead or Dummy-Residue—and Atomistic Models Based on Rigid-Body Refinement against SAS Data along with Specific Information Regarding the Uniqueness and Uncertainty of the Model and the Protocol Used to Obtain It

A prerequisite for archiving any model is the availability of the data specified in (1) and (2) so that the model can be critically evaluated against the original data and any subsequent data. Ab initio dummy-residue or bead-based shape models could be deposited as quasi-PDB files with, for example Cα atoms at bead positions but no sequence information. If generation of the bead model involved use of a derived $P(r)$ profile, then the $P(r)$ profile should be provided along with the parameters and program used to obtain it. If symmetry constraints were used for ab initio reconstructions, the results of analysis without symmetry constraints also should be presented to ensure that the anisometry of the symmetry-constrained model is correct. For models that utilize domains or subunits in a rigid-body refinement, the domains can be represented in the same manner as they entered the refinements, either as Cα-only or full-atom models with added glycans, heteroatoms, cofactors, and ligands. For models that are a combination of rigid bodies and beads, the representation can be a combination of the above.

All models should be accompanied by a detailed description of the protocol used to obtain it (including all parameters and software, with version numbers) along with evidence for the reproducibility of the reconstruction or rigid-body refinement and the existence of distinctly different solutions should be explored and results reported.

For ab initio bead or dummy-residue models, multiple reconstructions should have been performed, an assessment of the similarity of the resultant set of models provided, and, when appropriate, an average model deposited. For models developed using rigid-body refinement, consistency of multiple refinements should be demonstrated and any constraints used in the refinement (e.g., contacts, distances, orientation restrictions, etc.) must be documented in the deposition. If there are distinct classes of models that fit the SAS data equally well (ab initio or atomistic), at least one representative of each class should be included (e.g., using the available clustering tools; Petoukhov et al., 2012).

Criteria Need to Be Agreed on for Assessment of the Quality of SAS Data and Accuracy of SAS-Derived Models and the Extent to Which a Given Model Fits SAS Data

The quality of SAS data is critically dependent on the quality and intrinsic properties of the samples. For deriving reliable structural models from solution SAS measurements, evidence that the

solutions contain monodispersed, identical particles with a well-defined structure and no significant interparticle distance correlations must be provided. In this regard, a critical parameter to report is the molecular weight or volume of the scattering particle determined from the scattering data itself (from $I(0)$ analysis and/or concentration-independent methods based on the excluded (Porod) volume analysis). Bead models are essentially close-packed beads filling a volume such that the fit to the SAS data is optimized and there is no detailed stereochemistry. For atomistic models based on rigid-body refinement of crystallographic, NMR, or homology models against SAS data, the initial models are likely to have ill-defined stereochemistry at the linkages between domains and interfaces between domains or subunits. The domains or subunits themselves will be as accurate as the starting structures, but the linkers and interfaces are unlikely to be accurate at the atomic scale if they are being determined purely on the basis of SAS data. The real information in an atomistic model lies in the conformational torsion angles and these (together with assessment of any physically unrealistic "clashes") can be used to assess the quality of a model (through Ramachandran and rotamer analysis; Kleywegt, 2000, 2009; Kleywegt and Jones, 1995) and thus could contribute to an accurate and complete mapping of the uncertainty for a model. In considering the uncertainty in these models, it is important to note that rigid-body refined atomistic SAS-derived models cannot be expected to be uniformly reliable relative to their degrees of freedom; for example, center of mass separations are likely to be more accurate that rotation angles around long axes of objects with approximate cylindrical symmetry.

The common measure for the extent to which a model fits SAS data is a reduced χ^2, which is a global goodness-of-fit statistic of the theoretical model scattering to the measured data. Obtaining an ideal fit ($\chi^2 = 1.0$) depends on the assumption that the original measurements yield reliable counting statistics (which is not generally the case for image plate and CCD detectors, as they do not directly measure individual photon events) and that the propagated Poisson counting statistics fully account for the errors in the data. It may be that the absolute reduced χ^2 value is not relevant and one needs instead to demonstrate that a global minimum has been found. Also, as χ^2 is a global parameter, its absolute or minimum value also may be misleading and critical evaluation of the quality of the fit to the data requires inspection of the model fit over the entire measurement range. More robust nonparametric criteria can identify fits where systematic errors have been masked by the poor statistics. Comparison of the experimental data and the fit, as measured by the p value of a paired t test, allows one to test the goodness-of-fit without the use of experimental errors (Holm, 1979) and thus may be a preferred method.

Certain conditions must be fulfilled for the data to be considered sufficiently informative for model construction. These can be formulated based on the Shannon sampling theorem (Shannon and Weaver, 1949). The minimum q value must be smaller than the first Shannon channel ($q_{min} < \pi/D_{max}$) and it is suggested that that four to five channels are covered ($q_{max} > \sim$4–5 times π/D_{max}). There should be sufficient signal-to-noise ratio over the measured range to support the model. An average of no less than ten is suggested for a SAXS data set. For a SANS contrast series, the signal-to-noise requirement will be more variable as even low contrast and hence low signal-to-noise data sets can contribute to the overall solution.

With the Increasing Diversity of Structural Biology Data and Models Being Generated, Archiving Options for Models Derived from Diverse Data Will Be Required

Given the investment of resources required to develop models using diverse data, it is important that they are archived and available to the broader community in a form that permits evaluation, testing, and potential refinement. The scope of a future archive for SAS data and SAS-derived models could be broadened to include these kinds of models. Models based on diverse data must use a range of assumptions and the approaches to development of a particular model may be unique with different data types being given different weights. A complete description of the protocol used to develop the model should be provided so that it can be reproduced. These methods are not as well established as single data type based approaches, there is less experience with their accuracy, and consequently more apprehension concerning the validity of the resultant hybrid models. The crystallographic community has, through the work of wwPDB and its various task forces, developed standards and formats for data deposition and validation for specific kinds of data and associated models. These same criteria should be used in the evaluation of hybrid models, which should be accompanied with a complete map of uncertainty for all elements of the model (Lasker et al., 2012). Additional criteria may ultimately be required for hybrid models, beyond those that have been established for the single data type based models.

Thought Leaders from the Various Structural Biology Disciplines Should Jointly Define What to Archive in the PDB and What Complementary Archives Might Be Needed, Taking into Account Both Scientific Needs and Funding

The PDB is the global archive of biomacromolecular structure models that are atomistic and that have historically been expected to be reliable down to that level of detail, even though this is not always the case (e.g., when there is flexibility in the structure, the crystallographic data are low resolution, or the NMR ensemble indicates regions of high uncertainty). A broader conversation is needed to decide whether the PDB is the appropriate archive for SAS-derived and hybrid models where the measures of uncertainty are less well defined. The alternative is to have a separate archive that could be run in parallel with and tightly coupled to the PDB. This new "XDB" repository could be designed to fit and expose the strengths of the techniques and approaches used to produce the models, as opposed to forcing this distinct class of structural results to fit the requirements and expectations of atomistic models in the current PDB. The XDB would provide positive recognition of the increasing importance of models based on increasingly diverse data sets, from multiple heterogeneous sources, and incorporate the necessary flexibility for these kinds of results. Users would know that these models are distinct from the PDB structures, but they would be held with defined criteria for uniqueness and quality.

ACKNOWLEDGMENTS

The wwPDB SAS Task Force workshop that laid the foundations for this report was supported by members of the worldwide PDB: RCSB PDB (NSF DBI

0829586), PDBe (Wellcome Trust 088944), and PDBj (JST-NBDC). The research of J.A.T. on combining SAXS and NMR is supported in part by funding from Bruker.

REFERENCES

Alber, F., Förster, F., Korkin, D., Topf, M., and Sali, A. (2008). Integrating diverse data for structure determination of macromolecular assemblies. Annu. Rev. Biochem. *77*, 443–477.

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. *10*, 980.

Boehm, M.K., Woof, J.M., Kerr, M.A., and Perkins, S.J. (1999). The Fab and Fc fragments of IgA1 exhibit a different arrangement from that in IgG: a study by X-ray and neutron solution scattering and homology modelling. J. Mol. Biol. *286*, 1421–1447.

Christie, J.M., Arvai, A.S., Baxter, K.J., Heilmann, M., Pratt, A.J., O'Hara, A., Kelly, S.M., Hothorn, M., Smith, B.O., Hitomi, K., et al. (2012). Plant UVR8 photoreceptor senses UV-B by tryptophan-mediated disruption of cross-dimer salt bridges. Science *335*, 1492–1496.

Glatter, O., and Kratky, O. (1982). Small Angle X-ray Scattering (London: Academic Press.).

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

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

Guinier, A., and Fournet, G. (1955). Small-Angle Scattering of X-Rays (structure of matter series) (New York: Wiley).

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scand. J. Stat. *6*, 65–70.

Hura, G.L., Menon, A.L., Hammel, M., Rambo, R.P., Poole, F.L., 2nd, Tsutakawa, S.E., Jenney, F.E., Jr., Classen, S., Frankel, K.A., Hopkins, R.C., et al. (2009). Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). Nat. Methods *6*, 606–612.

Jacques, D.A., and Trewhella, J. (2010). Small-angle scattering for structural biology—expanding the frontier while avoiding the pitfalls. Protein Sci. *19*, 642–657.

Jacques, D.A., Langley, D.B., Jeffries, C.M., Cunningham, K.A., Burkholder, W.F., Guss, J.M., and Trewhella, J. (2008). Histidine kinase regulation by a cyclophilin-like inhibitor. J. Mol. Biol. *384*, 422–435.

Jacques, D.A., Langley, D.B., Hynson, R.M.G., Whitten, A.E., Kwan, A., Guss, J.M., and Trewhella, J. (2011). A novel structure of an antikinase and its inhibitor. J. Mol. Biol. *405*, 214–226.

Jacques, D.A., Guss, J.M., Svergun, D.I., and Trewhella, J. (2012). Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution. Acta Crystallogr. D Biol. Crystallogr. *68*, 620–626.

Karaca, E., and Bonvin, A.M. (2012). Advances in integrative modeling of biomolecular complexes. Methods.

Kleywegt, G.J. (2000). Validation of protein crystal structures. Acta Crystallogr. D Biol. Crystallogr. *56*, 249–265.

Kleywegt, G.J. (2009). On vital aid: the why, what and how of validation. Acta Crystallogr. D Biol. Crystallogr. *65*, 134–139.

Kleywegt, G.J., and Jones, T.A. (1995). Where freedom is given, liberties are taken. Structure *3*, 535–540.

Lasker, K., Förster, F., Bohn, S., Walzthoeni, T., Villa, E., Unverdorben, P., Beck, F., Aebersold, R., Sali, A., and Baumeister, W. (2012). Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. Proc. Natl. Acad. Sci. USA *109*, 1380–1387.

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W., et al. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. *487*, 545–574.

Malfois, M., and Svergun, D.I. (2000). sasCIF: an extension of core crystallographic information file for SAS. J. Appl. Cryst. *33*, 812–816.

Mertens, H.D., and Svergun, D.I. (2010). Structural characterization of proteins and complexes using small-angle X-ray solution scattering. J. Struct. Biol. *172*, 128–141. http://dx.doi.org/10.1016/j.jsb.2010.1006.1012.

Morgan, H.P., Schmidt, C.Q., Guariento, M., Blaum, B.S., Gillespie, D., Herbert, A.P., Kavanagh, D., Mertens, H.D.T., Svergun, D.I., Johansson, C.M., et al. (2011). Structural basis for engagement by complement factor H of C3b on a self surface. Nat. Struct. Mol. Biol. *18*, 463–470.

Nicastro, G., Todi, S.V., Karaca, E., Bonvin, A.M., Paulson, H.L., and Pastore, A. (2010). Understanding the role of the Josephin domain in the PolyUb binding and cleavage properties of ataxin-3. PLoS ONE *5*, e12430.

Nishimura, N., Hitomi, K., Arvai, A.S., Rambo, R.P., Hitomi, C., Cutler, S.R., Schroeder, J.I., and Getzoff, E.D. (2009). Structural mechanism of abscisic acid binding and signaling by dimeric PYR1. Science *326*, 1373–1379.

Petoukhov, M.V., Franke, D., Shkumatov, A.V., Tria, G., Kikhney, A.G., Gajda, M., Gorba, C., Mertens, H.D.T., Konarev, P.V., and Svergun, D.I. (2012). New developments in the ATSAS program package for small-angle scattering data analysis. J. Appl. Cryst. *45*, 342–350.

Putnam, C.D., Hammel, M., Hura, G.L., and Tainer, J.A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Q. Rev. Biophys. *40*, 191–285.

Rambo, R.P., and Tainer, J.A. (2010). Bridging the solution divide: comprehensive structural analyses of dynamic RNA, DNA, and protein assemblies by small-angle X-ray scattering. Curr. Opin. Struct. Biol. *20*, 128–137.

Rodrigues, J.P., Trellet, M., Schmitz, C., Kastritis, P., Karaca, E., Melquiond, A.S., and Bonvin, A.M. (2012). Clustering biomolecular complexes by residue contacts similarity. Proteins *80*, 1810–1817.

Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., and Sali, A. (2012). Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biol. *10*, e1001244.

Schiering, N., D'Arcy, A., Villard, F., Simic, O., Kamke, M., Monnet, G., Hassiepen, U., Svergun, D.I., Pulfer, R., Eder, J., et al. (2011). A macrocyclic HCV NS3/4A protease inhibitor interacts with protease and helicase residues in the complex with its full-length target. Proc. Natl. Acad. Sci. USA *108*, 21052–21056.

Schneidman-Duhovny, D., Rossi, A., Avila-Sakar, A., Kim, S.J., Velázquez-Muriel, J., Strop, P., Liang, H., Krukenberg, K.A., Liao, M., Kim, H.M., et al. (2012). A method for integrative structure determination of protein-protein complexes. Bioinformatics *28*, 3282–3289.

Shannon, C.E., and Weaver, W. (1949). The Mathematical Theory of Communication (Urbana: University of Illinois Press).

Whitten, A.E., Jeffries, C.M., Harris, S.P., and Trewhella, J. (2008). Cardiac myosin-binding protein C decorates F-actin: implications for cardiac function. Proc. Natl. Acad. Sci. USA *105*, 18360–18365.

Williams, R.S., Dodson, G.E., Limbo, O., Yamada, Y., Williams, J.S., Guenther, G., Classen, S., Glover, J.N.M., Iwasaki, H., Russell, P., and Tainer, J.A. (2009). Nbs1 flexibly tethers Ctp1 and Mre11-Rad50 to coordinate DNA double-strand break processing and repair. Cell *139*, 87–99.

CellPress

Structure
Meeting Review

Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop

Andrej Sali,[1,*] Helen M. Berman,[2] Torsten Schwede,[3] Jill Trewhella,[4] Gerard Kleywegt,[5] Stephen K. Burley,[2,6] John Markley,[7] Haruki Nakamura,[8] Paul Adams,[9,10] Alexandre M.J.J. Bonvin,[11] Wah Chiu,[12] Matteo Dal Peraro,[13] Frank Di Maio,[14] Thomas E. Ferrin,[15] Kay Grünewald,[16] Aleksandras Gutmanas,[5] Richard Henderson,[17] Gerhard Hummer,[18] Kenji Iwasaki,[19] Graham Johnson,[20] Catherine L. Lawson,[2] Jens Meiler,[21] Marc A. Marti-Renom,[22] Gaetano T. Montelione,[23,24] Michael Nilges,[25,26] Ruth Nussinov,[27,28] Ardan Patwardhan,[5] Juri Rappsilber,[29,30] Randy J. Read,[31] Helen Saibil,[32] Gunnar F. Schröder,[33,34] Charles D. Schwieters,[35] Claus A.M. Seidel,[36] Dmitri Svergun,[37] Maya Topf,[32] Eldon L. Ulrich,[7] Sameer Velankar,[5] and John D. Westbrook[2]

Structures of biomolecular systems are increasingly computed by integrative modeling that relies on varied types of experimental data and theoretical information. We describe here the proceedings and conclusions from the first wwPDB Hybrid/Integrative Methods Task Force Workshop held at the European Bioinformatics Institute in Hinxton, UK, on October 6 and 7, 2014. At the workshop, experts in various experimental fields of structural biology, experts in integrative modeling and visualization, and experts in data archiving addressed a series of questions central to the future of structural biology. How should integrative models be represented? How should the data and integrative models be validated? What data should be archived? How should the data and models be archived? What information should accompany the publication of integrative models?

Background

Historical Rationale for the Workshop

The PDB (http://wwpdb.org) was founded in 1971 with seven protein structures as its first holdings (Protein Data Bank, 1971). The global PDB archive now holds more than 100,000 atomic structures of biological macromolecules and their complexes, all of which are freely accessible. Most structures in the PDB archive (~90%) have been determined by X-ray crystallography, with the remainder contributed by two newer 3D structure determination methods, nuclear magnetic resonance (NMR) spectroscopy and 3D electron microscopy (3DEM).

Considerable effort has gone into understanding how to best curate the structural models and experimental data produced with these methods. Over the past several years, the Worldwide PDB (wwPDB; the global organization responsible for maintaining the PDB archive) (Berman et al., 2003) has established expert, method-specific task forces to advise on which experimental data and metadata from each method should be archived and how these data and the resulting structure models should be validated. The wwPDB X-ray Validation Task Force (VTF) made detailed recommendations on how to best validate structures determined by X-ray crystallography (Read et al., 2011). These

[1]Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, Byers Hall Room 503B, University of California, San Francisco, 1700 4th Street, San Francisco, CA 94158-2330, USA
[2]Research Collaboratory for Structural Bioinformatics Protein Data Bank, Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
[3]Swiss Institute of Bioinformatics Biozentrum, University of Basel, Klingelbergstrasse 50-70, 4056 Basel, Switzerland
[4]School of Molecular Bioscience, The University of Sydney, NSW 2006, Australia
[5]Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[6]Skaggs School of Pharmacy and Pharmaceutical Sciences and San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA
[7]BioMagResBank, Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706-1544, USA
[8]Protein Data Bank Japan, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan
[9]Physical Biosciences Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720-8235, USA
[10]Department of Bioengineering, UC Berkeley, Berkeley, CA 94720, USA
[11]Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Padualaan 8, Utrecht, 3584 CH, the Netherlands
[12]National Center for Macromolecular Imaging, Baylor College of Medicine, Houston, TX 77030, USA
[13]Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL) and Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[14]Department of Biochemistry, University of Washington, Seattle, WA 98195-7370, USA
[15]Department of Pharmaceutical Chemistry and Department of Bioengineering and Therapeutic Sciences, California Institute for Quantitative Biosciences, University of California, San Francisco, 600 16th Street, San Francisco, CA 94158-2517, USA
[16]Division of Structural Biology, Wellcome Trust Centre of Human Genetics, University of Oxford, OX3 7BN Oxford, UK

(Affiliations continued on next page)

recommendations have been implemented as a software pipeline used within the wwPDB Deposition and Annotation (D&A) system. Initial recommendations of the wwPDB NMR (Montelione et al., 2013) and Electron Microscopy (Henderson et al., 2012) VTFs have also been implemented. In addition, the wwPDB and, in later years, the Structural Biology Knowledgebase (SBKB), spearheaded three workshops focused on validation, archiving, and dissemination of comparative protein structure models (Berman et al., 2006; Schwede et al., 2009). It is anticipated that as new validation methods are developed and as more experience is gained with existing ones, additional validation procedures will be implemented in the wwPDB D&A system.

Increasingly, structures of very large macromolecular machines are being determined by combining observations from complementary experimental methods, including X-ray crystallography, NMR spectroscopy, 3DEM, small-angle scattering (SAS), crosslinking, and many others (Figure 1; Table 1). Data from these complementary methods are used to compute integrative or hybrid models (Ward et al., 2013). Atomic models produced in this fashion have been deposited in the PDB, but there is currently no mechanism within the PDB framework for archiving the experimental data generated by methods other than X-ray crystallography, NMR spectroscopy, and 3DEM. The most recently established task force, the wwPDB SAS Task Force (Trewhella et al., 2013), recommended creation of a SAS data and model repository that would interoperate with the PDB. The SAS Task Force also recommended that an international meeting be held to consider how best to deal with the archiving of data and models derived from integrative structure determination approaches.

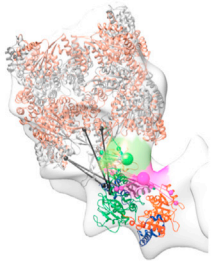
In response, a Hybrid/Integrative Methods Task Force was assembled by the wwPDB organization. Its inaugural meeting was held at the EMBL European Bioinformatics Institute (EBI) on October 6 and 7, 2014 (http://wwpdb.org/task/hybrid.php). In all, 38 participants from 37 academic and government institutions worldwide attended the workshop, which was co-chaired by Andrej Sali (University of California, San Francisco, USA), Torsten Schwede (Swiss Institute of Bioinformatics [SIB] and University of Basel, Switzerland), and Jill Trewhella (University of Sydney, Australia). Attendees included experts in relevant experimental techniques, integrative modeling, visualization, and data and model archiving.

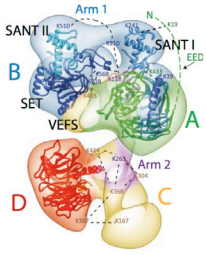
The workshop began with plenary talks followed by focused discussions. Gerard Kleywegt introduced the workshop objectives. Andrej Sali outlined the current state of integrative modeling. Helen Berman gave an overview of the history and status of the wwPDB organization. Jill Trewhella described the increasing role of SAS in integrative structural modeling, the need for the development of community standards and validation tools for biomolecular modeling using SAS data, and how SAS data and modeling resources could interoperate with the PDB. Claus Seidel outlined state-of-the-art single-molecule and ensemble Förster resonance energy transfer (FRET) spectroscopy (Kalinin et al., 2012) and live cell imaging, as well as related label-based spectroscopic methods for measuring select interatomic distances in macromolecular systems. Torsten Schwede presented the Protein Model Portal (Haas et al., 2013), including its linking of large databases of comparative models with experimental structure information in the PDB, and the Model Archive repository for all categories of *in silico* structural models.

Current Archives for Models and/or Supporting Data

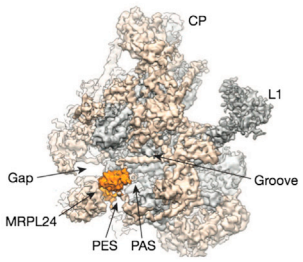
In this section, we review the PDB and management of data derived from crystallography, NMR spectroscopy, 3DEM, and

[17]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK
[18]Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Max-von-Laue Straße 3, 60438 Frankfurt am Main, Germany
[19]Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan
[20]Department of Bioengineering and Therapeutic Sciences, California Institute for Quantitative Biosciences, University of California, San Francisco, 600 16th Street, San Francisco, CA 94158-2330, USA
[21]Department of Chemistry, Center for Structural Biology, Vanderbilt University, Nashville, TN 37235, USA
[22]Genome Biology Group, Centre Nacional d'Anàlisi Genòmica (CNAG), Gene Regulation, Stem Cells and Cancer Program, Center for Genomic Regulation (CRG) and Institució Catalana de Recerca i Estudis Avançats (ICREA), 08028 Barcelona, Spain
[23]Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
[24]Department of Biochemistry, Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
[25]Département de Biologie Structurale et Chimie, Unité de Bioinformatique Structurale, Institut Pasteur, F-75015 Paris, France
[26]Unité Mixte de Recherche 3258, Centre National de la Recherche Scientifique, F-75015 Paris, France
[27]Cancer and Inflammation Program, Leidos Biomedical Research Inc., Frederick National Laboratory, National Cancer Institute, Frederick, MD 21702, USA
[28]Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel
[29]Wellcome Trust Centre for Cell Biology, Institute of Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, UK
[30]Department of Bioanalytics, Institute of Biotechnology, Technische Universität Berlin, 13355 Berlin, Germany
[31]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK
[32]Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck College, Malet Street, London WC1E 7HX, UK
[33]Institute of Complex Systems (ICS-6), Forschungszentrum Jülich, 52425 Jülich, Germany
[34]Physics Department, Heinrich-Heine University Düsseldorf, 40225 Düsseldorf, Germany
[35]Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, MD 20892-0520, USA
[36]Chair for Molecular Physical Chemistry, Heinrich-Heine-Universität, Universitätsstraße 1, 40225 Düsseldorf, Germany
[37]European Molecular Biology Laboratory, Hamburg Unit, Notkestrasse 85, 22607 Hamburg, Germany
*Correspondence: sali@salilab.org
http://dx.doi.org/10.1016/j.str.2015.05.013
All attendees of the Workshop are listed as authors.

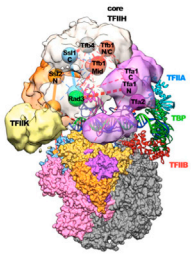

INO80



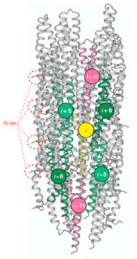
PRC2

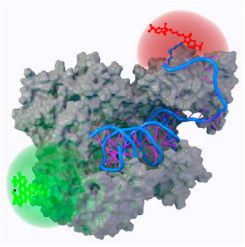


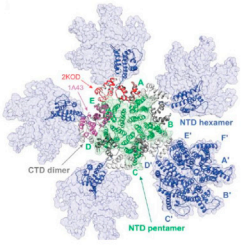
39S ribosome subunit



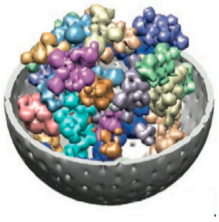
RNA pol II

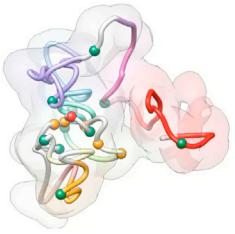

type III secretion system needle

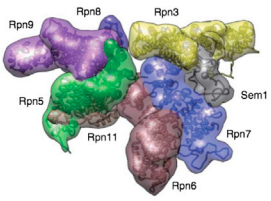

HIV-1 reverse transcriptase:DNA



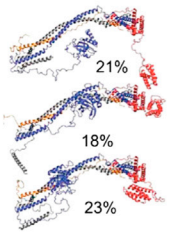
HIV capsid


genome architecture

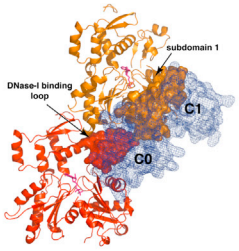

α-globin gene domain



proteosomal lid

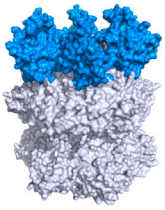


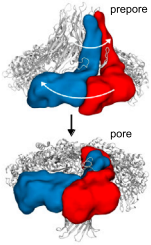
ESCRT-I complex



actin – myosin binding protein C

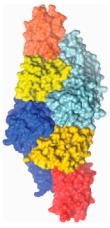

$[\Psi^{CD}]_2$

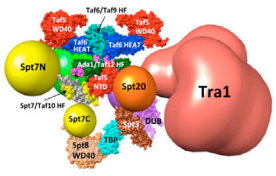

KaiB-KaiC complex



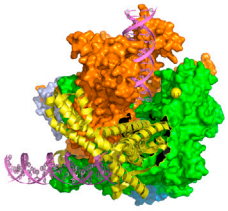
pore-forming toxin aerolysin

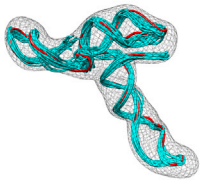

pleurotolysin

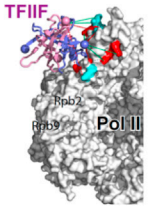

desulfurase-Isu-frataxin



SAGA transcriptional coactivator


bacterial RNA polymerase-promoter open complex


RNA ribosome-binding element



RNA polymerase II - transcription factor IIF

Figure 1. Examples of Recently Determined Integrative Structures

The molecular architecture of INO80 was determined with a 17-Å resolution cryo-electron microscopy (EM) map and 212 intra-protein and 116 inter-protein crosslinks (Russel et al., 2009). The molecular architecture of Polycomb Repressive Complex 2 (PRC2) was determined with a 21-Å resolution negative-stain EM map and ~60 intra-protein and inter-protein crosslinks (Shi et al., 2014). The molecular architecture of the large subunit of the mammalian mitochondrial ribosome (39S) was determined with a 4.9-Å resolution cryo-EM map and ~70 inter-protein crosslinks (Ward et al., 2013). The molecular architecture of the RNA polymerase II transcription pre-initiation complex was determined with a 16-Å resolution cryo-EM map plus 157 intra-protein and 109 inter-protein crosslinks (Alber et al., 2008). The atomic model of type III secretion system needle was determined with a 19.5-Å resolution cryo-EM map and solid-state nuclear magnetic resonance (NMR) data (Loquet et al., 2012). Molecular architecture of the productive HIV-1 reverse transcriptase:DNA primer-template complex in the open educt

(legend continued on next page)

Table 1. Types of Structural Data Used in Integrative Modeling

Structural Information	Method
Atomic structures of parts of the studied system	X-Ray and neutron crystallography, NMR spectroscopy, 3DEM, comparative modeling, and molecular docking
3D maps and 2D images	Electron microscopy and tomography
Atomic and protein distances	NMR, FRET, and other fluorescence techniques, DEER, EPR, and other spectroscopic techniques; chemical crosslinks detected by mass spectrometry, and disulfide bonds detected by gel electrophoresis
Binding site mapping	NMR spectroscopy, mutagenesis, FRET
Size, shape, and pairwise atomic distance distributions	SAS
Shape and size	Atomic force microscopy, ion mobility mass spectrometry, fluorescence correlation spectroscopy, and fluorescence anisotropy
Component positions	Super-resolution optical microscopy, FRET imaging
Physical proximity	Co-purification, native mass spectrometry, genetic methods, and gene/protein sequence covariance
Solvent accessibility	Footprinting methods, including H/D exchange assessed by mass spectrometry or NMR, and even functional consequences of point mutations
Proximity between different genome segments	Chromosome conformation capture and other data
Propensities for different interaction modes	Molecular mechanics force fields, potentials of mean force, statistical potentials, and sequence co-variation

Example methods that are informative about a variety of structural aspects of biomolecular systems are listed. 3DEM, 3D electron microscopy; DEER, double electron-electron resonance; EPR, electron paramagnetic resonance; FRET, Förster resonance energy transfer; H/D, hydrogen/deuterium; NMR, nuclear magnetic resonance; SAS, small-angle scattering.

SAS, plus archives for models derived exclusively on the basis on theoretical information.

PDB. For more than four decades, the PDB has served as the single global archive for atomic models of biological macromolecules; first for those derived from crystallography, and subsequently for models from NMR spectroscopy and 3DEM. The PDB also archives experimental data necessary to validate the structural models determined using these three methods. In addition, descriptions of the chemistry of polymers and ligands are collected, as are metadata describing sample preparation, experimental methods, model building, refinement statistics, literature references, and so forth. For all structural models in the PDB, geometric features are assessed with respect to standard valence geometry and intermolecular interactions, as recommended by the three wwPDB VTFs mentioned above.

Crystallography: Models and Data. For structures derived using X-ray, neutron, and combined X-ray/neutron crystallography, it has been mandatory to deposit structure factor amplitudes into the PDB since 2008 (http://www.wwpdb.org/news/news?year=2007#29-November-2007); until then, the submission of these primary data was optional. Additional validation against deposited structure factor amplitudes is carried out using procedures recommended by the X-ray VTF (Read et al., 2011). The resulting validation report includes graphical summaries of the quality of the overall model plus residue-specific features. Detailed assessments of various aspects of the model and its agreement with experimental and stereochemical data are also provided. In the near future, unmerged intensities will also be collected, enabling further validation activities.

state was determined by Förster resonance energy transfer (FRET) positioning and screening using a known HIV-1 reverse transcriptase structure (Kalinin et al., 2012). The structure of HIV-1 capsid protein was determined using residual dipolar couplings and small-angle X-ray scattering (SAXS) data (Deshmukh et al., 2013). The human genome architecture was determined based on tethered chromosome conformation capture and population-based modeling (Kalhor et al., 2012). The structural model of α-globin gene domain was determined based on Chromosome Conformation Capture Carbon Copy (5C) experiments (Bau et al., 2011). The molecular architecture of the proteosomal lid was determined using native mass spectrometry and 28 crosslinks (Politis et al., 2014). Structure models of the ESCRT-I complex were determined with SAXS, double electron-electron transfer, and FRET (Boura et al., 2011). Integrative model of actin and the cardiac myosin binding protein C was developed from a combination of crystallographic and NMR structures of subunits and domains, with positions and orientations optimized against SAXS and small-angle neutron scattering data to reveal information about the quaternary interactions (Whitten et al., 2008). The ensemble of $[\Psi^{CD}]_2$ NMR structures were fitted into the averaged cryo-electron tomography map (Miyazaki et al., 2010). Integrative model of the cyanobacterial circadian timing KaiB-KaiC complex was obtained based on hydrogen/deuterium exchange and collision cross-section data from mass spectrometry (Snijder et al., 2014). The pre-pore and pore conformations of the pore-forming toxin aerolysin were obtained combining cryo-EM data and molecular dynamics simulations (Degiacomi and Dal Peraro, 2013; Degiacomi et al., 2013). Segment of a pleurotolysin pore map (~11 Å resolution) with an ensemble of conformations shows the trajectory of β sheet opening during pore formation (Lukoyanova et al., 2015). A SAXS-based rigid-body model of a ternary complex of the iron-sulfur cluster assembly proteins desulfurase (orange) and scaffold protein Isu (blue) with bacterial ortholog of frataxin (yellow) was validated by NMR chemical shifts and mutagenesis (Prischi et al., 2010). The molecular architecture of the SAGA transcription coactivator complex was determined with 199 inter- and 240 intra-subunit crosslinks, several comparative models based on X-ray crystal structures, and a transcription factor IID core EM map at 31 Å resolution (Han et al., 2014). Structural organization of the bacterial (*Thermus aquaticus*) RNA polymerase-promoter open complex obtained by FRET (Mekler et al., 2002) was subsequently validated by a crystal structure (Zhang et al., 2012). The RNA ribosome-binding element from turnip crinkle virus genome was determined using NMR, SAXS, and EM data (Gong et al., 2015). The molecular architecture of the complex between RNA polymerase II and transcription factor IIF was determined using a deposited crystal structure of RNA polymerase II, homology models of some domains in transcription factor IIF, and 95 intra-protein and 129 inter-protein crosslinks (Chen et al., 2010).

NMR Spectroscopy: Models and Data. The Biological Magnetic Resonance DataBank (BioMagResBank or BMRB; http://www.bmrb.wisc.edu) is a repository for experimental and derived data gathered from NMR spectroscopic studies of biological molecules. The BMRB archive contains quantitative NMR spectral parameters, including assigned chemical shifts, coupling constants, and peak lists together with derived data, including relaxation parameters, residual dipolar couplings, hydrogen exchange rates, pK_a values, and so forth. Other data contained in the BMRB include: NMR restraints processed from original author depositions available from the PDB; time-domain spectral data from NMR experiments used to assign spectral resonances and determine structures of biological macromolecules; chemical shift and structure validation reports; and a database of 1D and 2D ^{1}H- and ^{13}C-NMR spectra for more than 1,200 metabolites. The BMRB website also provides tools for querying and retrieving data.

Since 2006, BMRB has been a member of the wwPDB organization (Markley et al., 2008). Chemical shift and restraint data that accompany model data are housed in both the BMRB and PDB archives. Deposited NMR data without model coordinates reside exclusively in the BMRB archive. The wwPDB D&A system provides for deposition, annotation, and validation of NMR models and related experimental data. Depositors of chemical shift and other data sets without accompanying models are automatically redirected to BMRB to deposit their data. Data exchange between the BRMB and PDB archives is facilitated by software tools utilizing correspondences maintained between the PDB Exchange Dictionary (PDBx) and the BMRB NMR-STAR Dictionary. Validation methods for NMR-derived models, measured chemical shifts, and restraint data are currently under development, in response to recommendations of the NMR VTF (Montelione et al., 2013). A working group composed of the major biomolecular NMR software developers has created a common NMR exchange format (NEF) for structural restraints, similar to NMR-STAR. The adoption of this NEF by NMR software developers will simplify data exchange and the archiving of NMR structural restraints by the wwPDB.

Electron Microscopy: Models and Maps. Atomistic structural models determined using 3DEM methods were first archived in the PDB in the 1990s. In 2002, the EM Data Bank (EMDB) was created by the Macromolecular Structure Database (now PDBe) at the EBI. In 2006, the EMDataBank (http://www.EMDataBank.org) was established as the unified global portal for one-stop deposition and retrieval of 3DEM density maps, atomic models, and associated metadata (Lawson et al., 2011). EMDataBank is a joint effort among PDBe, the Research Collaboratory for Structural Bioinformatics (RCSB) at Rutgers, and the National Center for Macromolecular Imaging (NCMI) at Baylor College of Medicine. EMDataBank also serves as a resource for news, events, software tools, data standards, raw data, and validation methods for the 3DEM community. 3DEM model and map data are now stored in separate branches of the wwPDB ftp archive site.

As for NMR-based models, the wwPDB D&A system supports processing of atomistic models and map data from 3DEM structure determinations. 3DEM map data deposited without atomistic models are stored exclusively in EMDB. Again, as for NMR, a mapping is maintained between the PDBx data dictionary and the EMDB XML-based data model. Validation methods for 3DEM maps and atomistic models are currently under development in response to recommendations from the EM VTF (Henderson et al., 2012).

SAS: Data and Model Archiving. The report from the first meeting of the wwPDB SAS Task Force (Trewhella et al., 2013) made the case for establishing "a global repository that holds standard format X-ray and neutron SAS data that is searchable and freely accessible for download" and that "options should be provided for including in the repository SAS-derived shape and atomistic models based on rigid-body refinement against SAS data along with specific information regarding the uniqueness and uncertainty of the model, and the protocol used to obtain it."

At present, there are two databases available for storing SAS data and models with associated metadata and analyses, both of which are freely accessible without limitations on data utilization via the Internet. As of March 2015, BIOISIS (http://www.bioisis.net/) contained 99 structures and is supported by teams at the Advanced Light Source and Diamond, while SASBDB (http://www.sasbdb.org/) (Valentini et al., 2015) contained 195 models and 114 experimental datasets and is supported by a team at EMBL-Hamburg.

Having evolved separately, these databases are distinctive in character. There was in principle agreement within the wwPDB SAS Task Force that BIOISIS and SASBDB will exchange data sets. Such exchange would be a step toward developing a federated approach to SAS data and model archiving, which in turn could ultimately be federated with the PDB, BMRB, and EMDB.

Further development of the sasCIF dictionary is required to permit full data exchange between the two SAS data repositories. sasCIF is a core crystallographic information file (CIF) developed to facilitate the SAS data exchange (Malfois and Svergun, 2000). As its name implies, sasCIF was implemented as an extension of the core CIF dictionary and has recently been extended to include new elements related to models, model fitting, validation tools, sample preparation, and experimental conditions (M.K., J.D.W., and D.S., unpublished data). sasCIFtools were developed as a documented set of publicly available programs for sasCIF data processing and format conversion; currently, SASBDB supports both import and export of sasCIF files.

Protein Model Portal. Comparative or homology modeling is routinely used to generate structural models of proteins for which experimentally determined structural models are not yet available (Marti-Renom et al., 2000; Schwede et al., 2009). Until 2006, such in silico models could be archived in the PDB, albeit in the absence of clear policies and procedures for their validation. Following recommendations from a stakeholder workshop convened in November 2005 (Berman et al., 2006), depositions to the PDB archive are limited to structural models substantially determined by experimental measurements from a defined physical sample (effective date October 15, 2006). The workshop also recommended that a central, publicly available archive or portal should be established for exclusively in silico models, and that methodology for estimating the accuracy of such computational models should be developed.

The Protein Model Portal (PMP) (Arnold et al., 2009; Haas et al., 2013) was developed at the SIB at the University of Basel

as a component of the SBKB (Berman et al., 2009; Gabanyi et al., 2011). Today, the SBKB integrates experimental information provided by the PDB with in silico models computed by automated modeling resources. In addition, the PMP provides access to several state-of-the-art model quality assessment services (Schwede et al., 2009). Since 2013, the Model Archive (http://modelarchive.org) resource has also served as a repository for individually generated in silico models of macromolecular structures, primarily those described in peer-reviewed publications. Finally, the Model Archive hosts all legacy models that were available from the PDB archive prior to 2006.

Each model in the PMP is assigned a stable, unique accession code (and digital object identifier or DOI) to ensure accurate cross-referencing in publications and other data repositories. Unlike experimentally determined structural models, in silico models are not the product of experimental measurements of a physical sample. They are generated computationally using various molecular modeling methods and underlying assumptions. Examples include comparative modeling, virtual docking of ligand molecules to protein targets, virtual docking of one protein to another, simulations of molecular dynamics and motions, and de novo (ab initio) protein modeling.

Effective archival storage of such models depends critically on capturing sufficient detail regarding underlying assumptions, parameters, methodology, and modeling constraints, to allow for assessment and faithful re-computation of the model. It is also essential that these models be accompanied by reliable estimates of uncertainty. In October 2013, a workshop on "Theoretical Model Archiving, Validation and PDBx/mmCIF Data Exchange Format" (http://www.proteinmodelportal.org/workshop-2013/) was hosted at Rutgers University to launch development of community standards for theoretical model archiving.

Integrative/Hybrid Structure Modeling

Motivation

Samples of many biological macromolecules prove recalcitrant to mainstream structural biology methods (i.e., crystallography, NMR, and 3DEM), because they are not crystallizable, are insoluble, are not of adequate purity, are conformationally heterogeneous, are too large or small, or do not remain intact during the course of the experiment. In such cases, integrative modeling is increasingly being used to compute structural models based on complementary experimental data and theoretical information (Figures 1 and 2; Table 1) (Alber et al., 2007, 2008; Robinson et al., 2007; Russel et al., 2012; Sali et al., 2003, 1990; Schneidman-Duhovny et al., 2014; Ward et al., 2013). Structural biology is no stranger to integrative models. Insights into the molecular details of the B-DNA double helix (Watson and Crick, 1953), the α helix, and the β sheet (Pauling et al., 1951) all depended on constructing structural models based on data derived from multiple sources (albeit without the benefit of digital computation). Integrative structure modeling of today has its origins in attempts to fit X-ray derived substructures into an EM density map of a larger assembly (Rayment et al., 1993). Other early examples include the model of the Gla-EGF domains from coagulation Factor X based on NMR and SAS data (Sunnerhagen et al., 1996), and the superhelical assembly of the bacteriophage fd gene 5 protein with single-stranded DNA based on neutron and X-ray SAS data, EM data, and the crystal structure of G5P (Olah et al., 1995); the latter study was inspired in part by molecular dynamics simulations guided by contacts from an NMR structure of the G5P dimer and EM data (Folmer et al., 1994).

Beyond overcoming sample limitations, the integrative approach has several additional advantages (Alber et al., 2007). First, synergy among the input data minimizes the drawbacks of sparse, noisy, and ambiguous data obtained from compositionally and structurally heterogeneous samples. Each individual piece of data may contain relatively little structural information, but by simultaneously fitting a model to all data derived from independent experiments, the uncertainty of the structures that fit the data can be markedly reduced. Second, the integrative approach can be used to produce all structural models consistent with available data, instead of myopically focusing on just one model. Third, comparison of an ensemble of structural models permits estimation of precision and, sometimes, the accuracy of both the experimental data and the model. Fourth, the integrative approach can make structural biologists more efficient by identifying which additional measurements are likely to have the greatest impact on integrative model precision and accuracy. Finally, integrative modeling provides a framework for considering perturbations of the system that are often required to collect the data; for example, spin labels are required for electron paramagnetic resonance experiments, membrane proteins are often reconstituted in micelles for NMR spectroscopy, and point mutations or even entire domains are introduced to stabilize preferred conformations for crystallization. While such perturbations complicate structural analysis, integrative modeling may allow us to distinguish biologically relevant states from artifacts of any individual approach. In summary, integrative structure determination maximizes the accuracy, precision, completeness, and efficiency of the structural coverage of biomolecular systems.

Experimental and Computational Methods for Generating Structural Information

Input information for integrative modeling can come from various experimental methods, physical theories, and statistical analyses of databases of known structures, biopolymer sequences, and interactions. These methods probe different structural aspects of the system (Table 1). In addition to information about average structures, numerous methods provide insights into dynamics of the system, which can also be incorporated into integrative modeling procedures (Russel et al., 2009). For example, both NMR spectroscopy and X-ray crystallography provide access to various measures of conformational dynamics; FRET, time-dependent double electron-electron resonance (DEER) spectroscopies, and even quantitative crosslinking/mass spectrometry (Fischer et al., 2013) can map distance changes in time; small-angle X-ray scattering (SAXS) can provide time-resolved information on the structures and processes with the temporal resolution of a millisecond; molecular dynamics simulations can map the dynamics of an atomic structure up to the millisecond timescale; and high-speed atomic force microscopy imaging can provide the dynamic live images of single molecules (Ando, 2014).

Approach

All structural characterization approaches correspond to finding models that best fit input information, as judged by use of

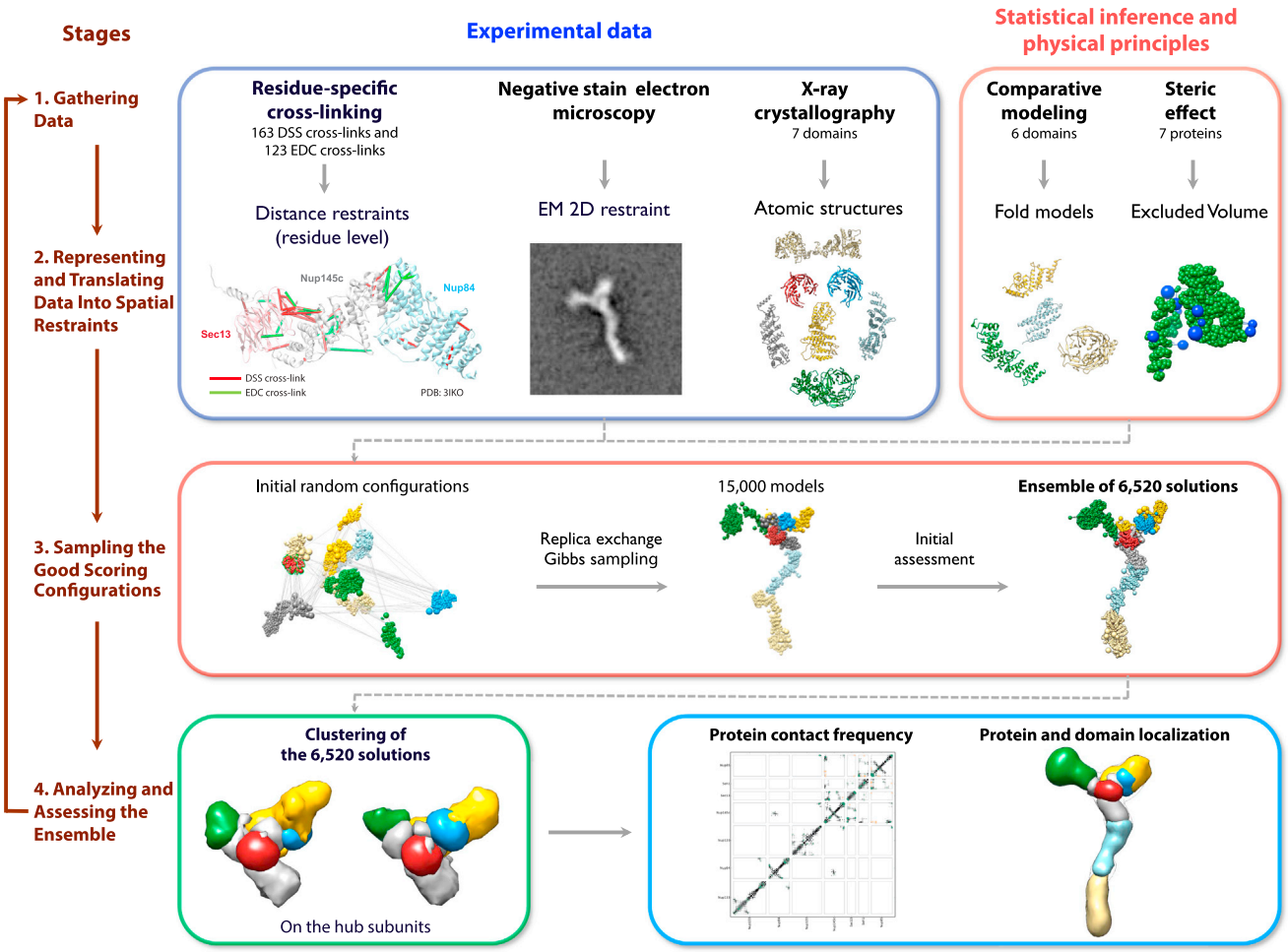


Figure 2. The Four Stages of Integrative Structure Determination
The approach is illustrated by its application to the heptameric Nup84 subcomplex of the nuclear pore complex (Shi et al., 2014).

a scoring function quantifying the difference between the observed data and the data computed from the model. Thus, any information about a structure determination target must always be converted to an explicit structural model through computation. Integrative approaches explicitly combine diverse experimental and theoretical information, with the goal of increasing accuracy, precision, coverage, and efficiency of structure determination. Input information can vary greatly in terms of resolution (i.e., precision, noise, uncertainty), accuracy, and quantity. All structure determination methods are integrative, albeit with differences in degree. At one end of the spectrum, even structure determination using predominantly crystallographic, NMR, or high-resolution single-particle EM data also generally requires a molecular mechanics force field description of atomic structure. At the other end of the spectrum, integrative methods rely more evenly on different types of information, often resulting in coarser models with higher uncertainty (Figure 1). Examples of such integrative methods include docking of comparative models of subunits into a 3DEM density map of the macromolecular assembly (Lasker et al., 2009); rigid-body fitting of multi-domain structures and complexes determined by crystallography or NMR to SAS data (Petoukhov and Svergun, 2005); and use of conformational sampling methods with sparse NMR data (Lange et al., 2012; Mueller et al., 2000), chemical crosslinks (Young et al., 2000), or even chemical shift data alone (Shen et al., 2008). It is not difficult to appreciate how integrative methods blur distinctions between models based primarily on theoretical considerations and those based primarily on experimental measurements from a physical sample.

The practice of integrative structure determination is iterative, consisting of four stages (Figure 2): gathering of data; choosing the representation and encoding of all data within a numerical scoring function consisting of spatial restraints; configurational sampling to identify structural models with good scores; and analyzing the models, including quantifying agreement with input spatial restraints and estimating model uncertainty. Input information about the system can be used to (1) select the set of variables that best represent the system (system representation), (2) rank the different configurations (scoring function), (3) search for good-scoring solutions (sampling); and (4) further filter good-scoring solutions produced by sampling.

Types of Integrative Models

A structural model of a macromolecular assembly is defined by the relative positions and orientations of its components

(e.g., atoms, pseudo-atoms, residues, secondary structure elements, domains, subunits, and subcomplexes). While traditional structural biology methods usually produce a single atomistic model, integrative models tend to be more complex in at least four respects. First, a model can be multi-scale (Grime and Voth, 2014), representing different levels of structural detail by a collection of geometrical primitives (e.g., points, spheres, tubes, 3D Gaussians, or probability densities). Thus, the same part of a system can be described with multiple representations and different parts of a system can be represented differently. An optimal representation facilitates accurate formulation of spatial restraints together with efficient and complete sampling of good-scoring solutions, while retaining sufficient detail (without over fitting) such that the resulting models are maximally useful for subsequent biological analysis (Schneidman-Duhovny et al., 2014). Second, a model can be multi-state, specifying multiple discrete states of the system required to explain the input information (each state may differ in structure, composition, or both) (Molnar et al., 2014; Pelikan et al., 2009). Third, a model can also specify the order of states in time and/or transitions between the states. This feature allows representation of a multi-step biological process, a functional cycle (Diez et al., 2004), a kinetic network (Pirchi et al., 2011), time evolution of a system (e.g., a molecular dynamics trajectory) (Bock et al., 2013), or FRET trajectories; for a comprehensive description of biomolecular function, it is essential to register state lifetimes, characteristic relaxation times, and direct rate constants. Finally, an ensemble of models may be provided to underscore the uncertainty in the input information, with each individual model satisfying the input information within an acceptable threshold (e.g., NMR-derived ensembles currently available in the PDB [Clore and Gronenborn, 1991; Snyder et al., 2005, 2014] and the ensembles generated from SAXS [Tria et al., 2015]). This aspect of the representation allows us to describe model uncertainty and to assess the completeness of input information; such ensembles are distinct from multiple states that represent actual variations in the structure, as implied by experimental information that cannot be accounted for by a single representative structure (Schneidman-Duhovny et al., 2014; Schröder, 2015).

Task Force Deliberations and Recommendations

Charge to the Task Force

A healthy debate is under way about how to classify structural models. A major motivation for this discussion is the lack of accurate general methods to assess the precision and accuracy of any model. As a result, models are often classified based on the predominant type of information used to compute them, which in turn tends to reflect the data-to-parameter ratio and thus model accuracy. However, as previously discussed, all structures are in fact integrative models that have been derived both from experimental measurements involving a physical sample of a biological macromolecule and prior knowledge of the underlying stereochemistry. It is therefore difficult, if not impossible, to draw definitive lines on the spectrum ranging from very well-determined ultra-high-resolution crystallographic structures (>40 experimental observations per non-hydrogen atom in the crystallographic asymmetric unit) and structural models based on a single or even no experimental observation.

Reflecting this debate about model classification, there are in principle several possibilities for archiving the models and associated data among distinct, publicly accessible model/data repositories, including: (1) a single mega-archive that serves as the repository for every type of structural model and data; (2) independent, free-standing repositories that house distinct types of models and data; and (3) a federated system of inter-operating repositories that archive models and data, with ''spheres of influence'' based on community consensus.

To address some of the challenges ahead and make recommendations about how best to proceed, the community stakeholders who assembled at the October 2014 meeting of the wwPDB Hybrid/Integrative Methods Task Force were divided into three discussion groups, each tasked with considering a series of related questions. What experimental data (beyond crystallography, NMR, and 3DEM) should be archived? Where and how should it be validated? What kinds of non-atomistic models can we expect and how should they be validated? What are the criteria for deciding where models should be archived? How should non-atomistic and mixed atomistic/non-atomistic models be archived? Should there be a separate archive for integrative (mixed) models (and data)? Should we establish a federated system of data and model archives to support integrative structural biology? The three breakout groups were asked to address these questions, report back with their findings, and make recommendations for the future. Each group independently approached the same set of questions. At the close of the meeting, the teams converged to compare notes, identify areas of commonality and diversity, and determine how best to move forward. The resulting consensus is reflected in this document.

Recommendations

Recommendation 1. In addition to archiving the models themselves, all relevant experimental data and metadata as well as experimental and computational protocols should be archived; inclusivity is key.

Ideally, structural models of any kind, derived by any method, should be archived.

Models are of greatest value when they are independently tested, potentially improved, and serve to further our understanding of how the function of a biological system is determined by its 3D structure(s). Therefore, models and necessary annotations must be freely available to the research community. The modeling process should be reproducible. Information concerning all aspects of a model should be deposited, including input data, corresponding spatial restraints, output models, and protocols used to convert input data into models. In addition to the input experimental data, the archival deposition should specify or include theoretically derived restraints used to compute the model (e.g., a statistical potential and a molecular mechanics force field). In practice, frequently used data types (e.g., distance information) should be prioritized for early complete implementation. Uncertainty in the input data needs to be well documented; some data uncertainty estimates may require modeling (e.g., Bayesian error estimates [Rieping et al., 2005]). Consistency between input data and the structural model should be documented as part of model validation.

Each expert community should drive decisions as to how much raw data, processed data, and metadata to deposit,

subject to the minimal requirement that the spatial restraints used for modeling must be derivable from the deposited information. Attention needs to be paid to annotating measurement conditions, such as temperature (Fenwick et al., 2014), sample concentration, environmental conditions (e.g., buffer), construct definition, and identification of all assembly components, all of which can significantly influence the experimental outcome. Cost-benefit analyses should be used to help guide which data should be archived. As much data as practical should be deposited, to facilitate model validation, future improvements of the model, and methods development (e.g., benchmarking sets). Of particular importance will be availability of some raw data to help drive improvement of data processing methods and for use by methods developers, who are often not generating the experimental data themselves.

Recommendation 2. A flexible model representation needs to be developed, allowing for multi-scale models, multi-state models, ensembles of models, and models related by time or other order.

Model representation should allow for as many types of "structural" models as possible, thereby encouraging collaboration among developers of integrative modeling software (Russel et al., 2012). At a minimum, the model representation should allow encoding of an ensemble of multi-scale, multi-state, time-ordered models (see the section on Types of Integrative Models). Uncertainty of the model coordinates should be tightly associated with the model coordinates in the model representation. Any model resident within an archive should be "self-contained" to facilitate utilization (e.g., for visualization). A common representation and format for models are useful for reasons of software interoperability. Particle-based representations/primitives need to be prioritized; non-particle-based model representations (e.g., continuum representations) merit further consideration by appropriate community stakeholders.

Recommendation 3. Procedures for estimating the uncertainty of integrative models should be developed, validated, and adopted.

Assessment of both an integrative model and the information on which it is based is of critical importance for guiding subsequent use of the model. For atomistic models, extant standard validation criteria from X-ray crystallography should be used. Beyond this test, validation of integrative models and data is a major research challenge that must be addressed and overcome. The following represent promising considerations (Alber et al., 2007; Schneidman-Duhovny et al., 2014): convergence of conformational sampling, fit of the model to the input information, test for clashes between geometrical primitives comprising the model, precision of the ensemble of solutions (visualized with, e.g., ribbon plots), cross-validation and statistical bootstrapping based on available data, tests based on data determined after the model was computed, and sensitivity analysis of the model to input data. Bayesian approaches may be particularly well suited to describe model uncertainty by computing posterior model densities from a forward model, noise model, and priors (Muschielok et al., 2008; Rieping et al., 2005). Tools for visualizing model validation should be developed.

Communities generating data used in integrative modeling should agree on the standard set of descriptors for data quality, as has been done for crystallography, NMR, and 3DEM.

Recommendation 4. A federated system of model and data archives should be created.

Integrative models can be based on a broad array of different experimental and computational techniques. While the specific spatial restraints implied by the data and used to construct an integrative model should be deposited with the model itself, the underlying experimental data often contain much richer information. This information should be captured in a federated system of domain-specific model and data archives. These individual member archives should be developed by community experts, based on method-specific standards for data archiving and validation. A federated system of model and data archives implies the need for a seamless exchange of information between independent archives. This seamless exchange requires a common dictionary of terms, agreed data formats, persistent and stable data object identifiers, and close synchronization of policies and procedures. Federated model and data archives need to develop efficient methods for data exchange to allow for transparent data access across the enterprise.

A single interface for the deposition of all data and models into the federated system is highly desirable. Such an interface would greatly facilitate the task of the depositor and, thereby, maximize compliance with deposition standards and requirements. In addition, reliance on a single entry point will help to ensure consistency across the federation at the time of deposition. Following successful deposition, individual datasets can be transferred to member databases for data curation and archiving if domain-specific databases exist. There should also be provision for collecting unstructured information in a "data commons," as proposed by the data science initiative at the NIH (Margolis et al., 2014).

Access to the contents of the federated database through a single portal is also most desirable, to facilitate dissemination of data, models, and experimental/computational protocols.

Of particular importance for integrative modeling will be the option to modify or update any aspect of the modeling procedure, for example, by adding new data. The federated archive should allow versioning for each deposited model. Such capabilities will facilitate the cycle of experiment and modeling, and accelerate production of more accurate, precise, and complete models (Russel et al., 2012).

Recommendation 5. Publication standards for integrative models should be established.

Over the past decade, the wwPDB organization has worked with relevant scientific journals to help establish publication standards for structural models coming from crystallography, NMR spectroscopy, and 3DEM. Community standards now include requiring authors to make their validation reports available to reviewers and editors. Through the International Union of Crystallography Small Angle Scattering and Journals Commissions, the SAS community developed and agreed upon publication guidelines for structural modeling of biomolecules therefrom (Jacques et al., 2012). A set of standards for publishing integrative models should be developed along similar lines.

Implementation

Implementation of Recommendation 1 poses a host of cultural and technical challenges. Experimentalists and modelers need to provide the data, models, and protocols, thus at least partly addressing increasing concerns regarding reproducibility of

scientific results. From a technical perspective, inter-operating data dictionaries for all methods need to be created. In addition, potential storage bottlenecks need to be addressed.

Implementation of Recommendations 2 and 3 will require significant research as to how best to represent and validate the many different kinds of integrative models. In addition, the community will need to agree on a common set of standards that are sufficiently mutable to allow for future innovation. Efforts such as the ''Cryo-EM Modeling Challenge'' may facilitate this process (http://www.emdatabank.org/modeling_chllnge).

Implementation of Recommendation 4 will require agreement on a common data exchange system among member repositories. Based on past accomplishments, the wwPDB is well positioned to play a leadership role in establishing the proposed federated system, including provision of common deposition and access interfaces. The wwPDB should begin this process by providing training and advice on data archiving and curation to contributing domain-specific member repositories.

Implementation of Recommendation 5 will require continued work with the journals that publish structural models of biological macromolecules.

Significant resources will be required to implement these recommendations, including grants for research, infrastructure, and workshops. These efforts are international by their very nature and will require funding from multiple public and private sources, including in North America, Europe, and Asia.

ACKNOWLEDGMENTS

The workshop was supported by funding to PDBe by Wellcome Trust 088944; RCSB PDB by NSF DBI 1338415; PDBj by JST-NBDC; BMRB by NLM P41 LM05799; EMDataBank by NIH GM079429; and tax-deductible donations made to the wwPDB Foundation in support of wwPDB outreach activities.

REFERENCES

Alber, F., Dokudovskaya, S., Veenhoff, L., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B., et al. (2007). Determining the architectures of macromolecular assemblies. Nature *450*, 683–694.

Alber, F., Förster, F., Korkin, D., Topf, M., and Sali, A. (2008). Integrating diverse data for structure determination of macromolecular assemblies. Annu. Rev. Biochem. *77*, 443–477.

Ando, T. (2014). High-speed AFM imaging. Curr. Opin. Struct. Biol. *28*, 63–68.

Arnold, K., Kiefer, F., Kopp, J., Battey, J.N., Podvinec, M., Westbrook, J.D., Berman, H.M., Bordoli, L., and Schwede, T. (2009). The Protein Model Portal. J. Struct. Funct. Genomics *10*, 1–8.

Bau, D., Sanyal, A., Lajoie, B.R., Capriotti, E., Byron, M., Lawrence, J.B., Dekker, J., and Marti-Renom, M.A. (2011). The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. Nat. Struct. Mol. Biol. *18*, 107–114.

Berman, H.M., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. *10*, 980.

Berman, H.M., Burley, S.K., Chiu, W., Sali, A., Adzhubei, A., Bourne, P.E., Bryant, S.H., Dunbrack, R.L., Jr., Fidelis, K., Frank, J., et al. (2006). Outcome of a workshop on archiving structural models of biological macromolecules. Structure *14*, 1211–1217.

Berman, H.M., Westbrook, J.D., Gabanyi, M.J., Tao, W., Shah, R., Kouranov, A., Schwede, T., Arnold, K., Kiefer, F., Bordoli, L., et al. (2009). The protein structure initiative structural genomics knowledgebase. Nucleic Acids Res. *37*, D365–D368.

Bock, L.V., Blau, C., Schröder, G.F., Davydov, I.I., Fischer, N., Stark, H., Rodnina, M.V., Vaiana, A.C., and Grubmuller, H. (2013). Energy barriers and driving forces in tRNA translocation through the ribosome. Nat. Struct. Mol. Biol. *20*, 1390–1396.

Boura, E., Rozycki, B., Herrick, D.Z., Chung, H.S., Vecer, J., Eaton, W.A., Cafiso, D.S., Hummer, G., and Hurley, J.H. (2011). Solution structure of the ESCRT-I complex by small-angle X-ray scattering, EPR, and FRET spectroscopy. Proc. Natl. Acad. Sci. USA *108*, 9437–9442.

Chen, Z.A., Jawhari, A., Fischer, L., Buchen, C., Tahir, S., Kamenski, T., Rasmussen, M., Lariviere, L., Bukowski-Wills, J.C., Nilges, M., et al. (2010). Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. EMBO J. *29*, 717–726.

Clore, G.M., and Gronenborn, A.M. (1991). Structures of larger proteins in solution: three- and four-dimensional heteronuclear NMR spectroscopy. Science *252*, 1390–1399.

Degiacomi, M.T., and Dal Peraro, M. (2013). Macromolecular symmetric assembly prediction using swarm intelligence dynamic modeling. Structure *21*, 1097–1106.

Degiacomi, M.T., Iacovache, I., Pernot, L., Chami, M., Kudryashev, M., Stahlberg, H., van der Goot, F.G., and Dal Peraro, M. (2013). Molecular assembly of the aerolysin pore reveals a swirling membrane-insertion mechanism. Nat. Chem. Biol. *9*, 623–629.

Deshmukh, L., Schwieters, C.D., Grishaev, A., Ghirlando, R., Baber, J.L., and Clore, G.M. (2013). Structure and dynamics of full-length HIV-1 capsid protein in solution. J. Am. Chem. Soc. *135*, 16133–16147.

Diez, M., Zimmermann, B., Börsch, M., König, M., Schweinberger, E., Steigmiller, S., Reuter, R., Felekyan, S., Kudryavtsev, V., Seidel, C.A.M., and Gräber, P. (2004). Proton-powered subunit rotation in single membrane-bound F0 F1-ATP synthase. Nat. Struct. Mol. Biol. *11*, 135–141.

Fenwick, R.B., van den Bedem, H., Fraser, J.S., and Wright, P.E. (2014). Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. Proc. Natl. Acad. Sci. USA *111*, E445–E454.

Fischer, L., Chen, Z.A., and Rappsilber, J. (2013). Quantitative cross-linking/mass spectrometry using isotope-labelled cross-linkers. J. Proteomics *88*, 120–128.

Folmer, R.H., Nilges, M., Folkers, P.J., Konings, R.N., and Hilbers, C.W. (1994). A model of the complex between single-stranded DNA and the single-stranded DNA binding protein encoded by gene V of filamentous bacteriophage M13. J. Mol. Biol. *240*, 341–357.

Gabanyi, M.J., Adams, P.D., Arnold, K., Bordoli, L., Carter, L.G., Flippen-Andersen, J., Gifford, L., Haas, J., Kouranov, A., McLaughlin, W.A., et al. (2011). The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. J. Struct. Funct. Genomics *12*, 45–54.

Gong, Z., Schwieters, C.D., and Tang, C. (2015). Conjoined use of EM and NMR in RNA structure refinement. PLoS One *10*, e0120445.

Grime, J.M.A., and Voth, G.A. (2014). Highly scalable and memory efficient ultra-coarse-grained molecular dynamics simulations. J. Chem. Theory Comp. *10*, 423–431.

Haas, J., Roth, S., Arnold, K., Kiefer, F., Schmidt, T., Bordoli, L., and Schwede, T. (2013). The Protein Model Portal—a comprehensive resource for protein structure and model information. Database *2013*, bat031.

Han, Y., Luo, J., Ranish, J., and Hahn, S. (2014). Architecture of the *Saccharomyces cerevisiae* SAGA transcription coactivator complex. EMBO J. *33*, 2534–2546.

Henderson, R., Sali, A., Baker, M.L., Carragher, B., Devkota, B., Downing, K.H., Egelman, E.H., Feng, Z., Frank, J., Grigorieff, N., et al. (2012). Outcome of the first Electron Microscopy Validation Task Force meeting. Structure *20*, 205–214.

Jacques, D.A., Guss, J.M., Svergun, D.I., and Trewhella, J. (2012). Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution. Acta Crystallogr. D Biol. Crystallogr. *68*, 620–626.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotechnol. *30*, 90–98.

Kalinin, S., Peulen, T., Sindbert, S., Rothwell, P.J., Berger, S., Restle, T., Goody, R.S., Gohlke, H., and Seidel, C.A.M. (2012). A toolkit and benchmark study for FRET-restrained high-precision structural modeling. Nat. Methods *9*, 1218–1225.

Lange, O.F., Rossi, P., Sgourakis, N.G., Song, Y., Lee, H.W., Aramini, J.M., Ertekin, A., Xiao, R., Acton, T.B., Montelione, G.T., and Baker, D. (2012). Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proc. Natl. Acad. Sci. USA *109*, 10873–10878.

Lasker, K., Topf, M., Sali, A., and Wolfson, H.J. (2009). Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. J. Mol. Biol. *388*, 180–194.

Lawson, C.L., Baker, M.L., Best, C., Bi, C., Dougherty, M., Feng, P., van Ginkel, G., Devkota, B., Lagerstedt, I., Ludtke, S.J., et al. (2011). EMDataBank.org: unified data resource for CryoEM. Nucleic Acids Res. *39*, D456–D464.

Loquet, A., Sgourakis, N.G., Gupta, R., Giller, K., Riedel, D., Goosmann, C., Griesinger, C., Kolbe, M., Baker, D., Becker, S., and Lange, A. (2012). Atomic model of the type III secretion system needle. Nature *486*, 276–279.

Lukoyanova, N., Kondos, S.C., Farabella, I., Law, R.H., Reboul, C.F., Caradoc-Davies, T.T., Spicer, B.A., Kleifeld, O., Traore, D.A., Ekkel, S.M., et al. (2015). Conformational changes during pore formation by the perforin-related protein pleurotolysin. PLoS Biol. *13*, e1002049.

Malfois, M., and Svergun, D. (2000). sasCIF: an extension of core crystallographic information file for SAS. J. App. Cryst. *33*, 812–816.

Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., Guyer, M., and Green, E.D. (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J. Am. Med. Inform. Assoc. *21*, 957–958.

Markley, J.L., Ulrich, E.L., Berman, H.M., Henrick, K., Nakamura, H., and Akutsu, H. (2008). BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. J. Biomol. NMR *40*, 153–155.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. *29*, 291–325.

Mekler, V., Kortkhonjia, E., Mukhopadhyay, J., Knight, J., Revyakin, A., Kapanidis, A.N., Niu, W., Ebright, Y.W., Levy, R., and Ebright, R.H. (2002). Structural organization of bacterial RNA polymerase holoenzyme and the RNA polymerase-promoter open complex. Cell *108*, 599–614.

Miyazaki, Y., Irobalieva, R.N., Tolbert, B.S., Smalls-Mantey, A., Iyalla, K., Loeliger, K., D'Souza, V., Khant, H., Schmid, M.F., Garcia, E.L., et al. (2010). Structure of a conserved retroviral RNA packaging element by NMR spectroscopy and cryo-electron tomography. J. Mol. Biol. *404*, 751–772.

Molnar, K.S., Bonomi, M., Pellarin, R., Clinthorne, G.D., Gonzalez, G., Goldberg, S.D., Goulian, M., Sali, A., and DeGrado, W.F. (2014). Cys-scanning disulfide crosslinking and bayesian modeling probe the transmembrane signaling mechanism of the histidine kinase, PhoQ. Structure *22*, 1239–1251.

Montelione, G.T., Nilges, M., Bax, A., Güntert, P., Herrmann, T., Markley, J.L., Richardson, J., Schwieters, C., Vuister, G.W., Vranken, W., and Wishart, D. (2013). Recommendations of the wwPDB NMR Structure Validation Task Force. Structure *21*, 1563–1570.

Mueller, G.A., Choy, W.Y., Yang, D., Forman-Kay, J.D., Venters, R.A., and Kay, L.E. (2000). Global folds of proteins with low densities of NOEs using residual dipolar couplings: application to the 370-residue maltodextrin-binding protein. J. Mol. Biol. *300*, 197–212.

Muschielok, A., Andrecka, J., Jawhari, A., Bruckner, F., Cramer, P., and Michaelis, J. (2008). A nano-positioning system for macromolecular structural analysis. Nat. Methods *5*, 965–971.

Olah, G.A., Gray, D.M., Gray, C.W., Kergil, D.L., Sosnick, T.R., Mark, B.L., Vaughan, M.R., and Trewhella, J. (1995). Structures of fd gene 5 protein.nucleic acid complexes: a combined solution scattering and electron microscopy study. J. Mol. Biol. *249*, 576–594.

Pauling, L., Corey, R.B., and Branson, H.R. (1951). The structures of proteins. Proc. Natl. Acad. Sci. USA *37*, 205.

Pelikan, M., Hura, G.L., and Hammel, M. (2009). Structure and flexibility within proteins as identified through small angle X-ray scattering. Gen. Physiol. Biophys. *28*, 174–189.

Petoukhov, M.V., and Svergun, D.I. (2005). Global rigid body modeling of macromolecular complexes against small-angle scattering data. Biophys. J. *89*, 1237–1250.

Pirchi, M., Ziv, G., Riven, I., Cohen, S.S., Zohar, N., Barak, Y., and Haran, G. (2011). Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein. Nat. Commun. *2*, 493.

Politis, A., Stengel, F., Hall, Z., Hernandez, H., Leitner, A., Walzthoeni, T., Robinson, C.V., and Aebersold, R. (2014). A mass spectrometry-based hybrid method for structural modeling of protein complexes. Nat. Methods *11*, 403–406.

Prischi, F., Konarev, P.V., Iannuzzi, C., Pastore, C., Adinolfi, S., Martin, S.R., Svergun, D.I., and Pastore, A. (2010). Structural bases for the interaction of frataxin with the central components of iron-sulphur cluster assembly. Nat. Commun. *1*, 95.

Protein Data Bank. (1971). Protein Data Bank. Nat. New Biol. *233*, 223.

Rayment, I., Holden, H.M., Whittaker, M., Yohn, C.B., Lorenz, M., Holmes, K.C., and Milligan, R.A. (1993). Structure of the actin-myosin complex and its implications for muscle contraction. Science *261*, 58–65.

Read, R.J., Adams, P.D., Arendall, W.B., III, Brunger, A.T., Emsley, P., Joosten, R.P., Kleywegt, G.J., Krissinel, E.B., Lutteke, T., Otwinowski, Z., et al. (2011). A new generation of crystallographic validation tools for the Protein Data Bank. Structure *19*, 1395–1412.

Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential structure determination. Science *309*, 303–306.

Robinson, C.V., Sali, A., and Baumeister, W. (2007). The molecular sociology of the cell. Nature *450*, 973–982.

Russel, D., Lasker, K., Phillips, J., Schneidman-Duhovny, D., Velazquez-Muriel, J., and Sali, A. (2009). The structural dynamics of macromolecular processes. Curr. Opin. Cell Biol. *21*, 97–108.

Russel, D., Lasker, K., Webb, B., Velazquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., and Sali, A. (2012). Putting the pieces together: integrative structure determination of macromolecular assemblies. PLoS Biol. *10*, e1001244.

Sali, A., Overington, J.P., Johnson, M.S., and Blundell, T.L. (1990). From comparisons of protein sequences and structures to protein modelling and design. Trends Biochem. Sci. *15*, 235–240.

Sali, A., Glaeser, R., Earnest, T., and Baumeister, W. (2003). From words to literature in structural proteomics. Nature *422*, 216–225.

Schneidman-Duhovny, D., Pellarin, R., and Sali, A. (2014). Uncertainty in integrative structural modeling. Curr. Opin. Struct. Biol. *28*, 96–104.

Schröder, G.F. (2015). Hybrid methods for macromolecular structure determination: experiment with expectations. Curr. Opin. Struct. Biol. *31*, 20–27.

Schwede, T., Sali, A., Honig, B., Levitt, M., Berman, H.M., Jones, D., Brenner, S.E., Burley, S.K., Das, R., Dokholyan, N.V., et al. (2009). Outcome of a workshop on applications of protein models in biomedical research. Structure *17*, 151–159.

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J.M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K.K., Lemak, A., et al. (2008). Consistent blind protein structure generation from NMR chemical shift data. Proc. Natl. Acad. Sci. USA *105*, 4685–4690.

Shi, Y., Fernandez-Martinez, J., Tjioe, E., Pellarin, R., Kim, S.J., Williams, R., Schneidman, D., Sali, A., Rout, M.P., and Chait, B.T. (2014). Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. Mol. Cell. Proteomics *13*, 2927–2943.

Snijder, J., Burnley, R.J., Wiegard, A., Melquiond, A.S.J., Bonvin, A.M.J.J., Axmann, I.M., and Heck, A.J.R. (2014). Insight into cyanobacterial circadian timing from structural details of the KaiB-KaiC interaction. Proc. Natl. Acad. Sci. USA *111*, 1379–1383.

Snyder, D.A., Bhattacharya, A., Huang, Y.J., and Montelione, G.T. (2005). Assessing precision and accuracy of protein structures derived from NMR data. Proteins *59*, 655–661.

Snyder, D.A., Grullon, J., Huang, Y.J., Tejero, R., and Montelione, G.T. (2014). The expanded FindCore method for identification of a core atom set for assessment of protein structure prediction. Proteins *82* (*Suppl 2*), 219–230.

Sunnerhagen, M., Olah, G.A., Stenflo, J., Forsen, S., Drakenberg, T., and Trewhella, J. (1996). The relative orientation of Gla and EGF domains in coagulation factor X is altered by Ca^{2+} binding to the first EGF domain. A combined NMR-small angle X-ray scattering study. Biochemistry *35*, 11547–11559.

Trewhella, J., Hendrickson, W.A., Kleywegt, G.J., Sali, A., Sato, M., Schwede, T., Svergun, D.I., Tainer, J.A., Westbrook, J., and Berman, H.M. (2013). Report of the wwPDB Small-Angle Scattering Task Force: data requirements for biomolecular modeling and the PDB. Structure *21*, 875–881.

Tria, G., Mertens, H.D.T., Kachala, M., and Svergun, D.I. (2015). Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. IUCrJ *2*, 207–217.

Valentini, E., Kikhney, A.G., Previtali, G., Jeffries, C.M., and Svergun, D.I. (2015). SASBDB, a repository for biological small-angle scattering data. Nucleic Acids Res. *43*, D357–D363.

Ward, A., Sali, A., and Wilson, I. (2013). Integrative structural biology. Science *339*, 913–915.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737–738.

Whitten, A.E., Jeffries, C.M., Harris, S.P., and Trewhella, J. (2008). Cardiac myosin-binding protein C decorates F-actin: implications for cardiac function. Proc. Natl. Acad. Sci. USA *105*, 18360–18365.

Young, M.M., Tang, N., Hempel, J.C., Oshiro, C.M., Taylor, E.W., Kuntz, I.D., Gibson, B.W., and Dollinger, G. (2000). High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. Proc. Natl. Acad. Sci. USA *97*, 5802–5806.

Zhang, Y., Feng, Y., Chatterjee, S., Tuske, S., Ho, M.X., Arnold, E., and Ebright, R.H. (2012). Structural basis of transcription initiation. Science *338*, 1076–1080.

Jill Trewhella
Professor Emeritus

03 February 2017

Small-Angle Scattering Validation Task Force Members
Wayne A. Hendrickson, Columbia University
Andrej Sali, UC San Francisco
Torsten Schwede, Biozentrum, University of Basel
Dmitri I. Svergun, EMBL Hamburg
John A. Tainer, U Texas MD Anderson Cancer Center

Transmitted: by email

Dear Colleagues

I am contacting you to request feedback on membership of our task force and to propose a virtual meeting of the Small-Angle Scattering validation task force (SASvtf) with the following agenda:

1. review progress to date on the implementation of our recommendations as outlined in our 2013 report in Structure and plan next steps; and
2. consider in detail the fourth recommendation from the SASvtf; that is "criteria need to be agreed upon for the assessment of the quality of deposited data and the accuracy of SAS-derived models, and the extent to which a given model fits the SAS data."

With respect to item 2, in our 2013 report (Trewhella et al, 2013, attached) we expanded upon preliminary guidelines (Jacques et al., 2012, attached) that addressed **sample quality**, **data acquisition and reduction**, **presentation of scattering data and validation**, and **modelling** for biomolecular SAS. These preliminary guidelines were developed in consultation with the IUCr SAS Commission and other experts in the field and formally adopted by the IUCr Journals. They have subsequently been used by a growing cohort of SAS researchers.

I am in the process of drafting an update of the preliminary publication guidelines to a v1.0 that would incorporate the additional points made in from our 2013 report, as well as input from the IUCr SAS Commission and other experts in the field. I propose to circulate this draft to the SASvtf for comment. At our virtual meeting we would review the comments and agree a final set of guidelines, including specific, implementable recommendations for both data and model validation, which we would formally recommend as the SASvtf.

I plan to present the revised, updated publication guidelines at the IUCr Congress in Hyderabad in August during an open meeting of the IUCr SAS Commission that I have

Biochemistry Bldg, G08
Faculty of Science
The University of Sydney
NSW 2006 Australia

T +61 (0)4 3275 3933
E jill.trewhella@sydney.edu.au
http://sydney.edu.au/science/people/jill.trewhella.php

ABN 15 211 513 464
CRICOS 00026A

called for as Chair of the SAS Commission. Subsequently we will submit the paper for publication describing the revsied guidelines and the basis for recommended adjustments.

As we embark on this next phase of work, I propose expanding the SAS expertise on our task force and would be happy to entertain any and all suggestions for additions. I would like to suggest that we recruit from among Masaaki Sugiyama (Kyoto), Patrice Vachette (Saclay Paris), Frank Gabel (Institut de Biologie Structurale, Genoble), Lois Pollack (Cornell), and Dina Schneidman (Hebrew University) each of whom is publishing high quality work using SAS. With these additions, we can diversify and strengthen our expertise with respect to SANS, combined SANS/NMR, SAXS on large virus particles and RNA.

Please let me know your thoughts on membership and I will be asking the RSCB PDB to assist in setting up the virtual meeting in early May at a time that best suits all participants.

Yours sincerely,

Jill Trewhella, PhD

cc John L. Markley, BMRB
Haruki Nakamura, PDBj
Sameer Valenkar, PDBe
Stephen K. Burley, PDB RSCB
Helen Berman, PDB RSCB

Jill Trewhella
Professor Emeritus

16 February 2017

wwPDB Leadership Group
John L. Markley, BMRB
Haruki Nakamura, PDBj
Sameer Valenkar, PDBe
Stephen K. Burley, PDB RSCB
Helen Berman, PDB RSCB

Dear Colleagues

I write to you with an update in my capacity as Chair of the Small-Angle Scattering validation task force (SASvtf). In my letter of 2 February to the current membership of that committee, I noted that one of the six key recommendations from our 2013 report was that "criteria need to be agreed upon for the assessment of the quality of deposited data and the accuracy of SAS-derived models, and the extent to which a given model fits the SAS data." That report further built on the publication guidelines for biomolecular SAS published in *Acta D* in 2012 that was the result of a community effort led by the IUCr Commission for SAS (CSAS). A priority piece of work for the SASvtf now is to update the criteria given advances since these earlier publications and to provide advice to the wwPDB on "specific, implementable recommendations for both data and model validation."

To accomplish this next phase, the SASvtf needs greater breadth of SAS expertise. The committee currently includes only myself and Dmitri Svergun and John Tainer as SAS specialists. Additionally, as there is significant overlap between the work of the SASvtf and the IUCr Commission on SAS (CSAS) in data and model validation, it is desirable that the two efforts do not put out conflicting recommendations.

I therefore propose to expand the current membership of the SASvtf, first to broaden the expertise and second to ensure some overlap with the IUCr CSAS membership and international engagement. The most important criterion however was to choose active researchers publishing the highest quality biomolecular SAS work. Unless I hear alternative suggestions for consideration, I plan to invite the following individuals to join the SASvtf.

Masaaki Sugiyama (Kyoto) – is Japan's national committee recommendation for the IUCr for CSAS membership, endorsed by the current CSAS for appointment August 2017.
https://www.labome.org/expert/japan/kyoto/sugiyama/masaaki-sugiyama-1208789.html

Patrice Vachette (Institute Pasteur, Saclay Paris) – a long standing and respected expert in SAXS, both laboratory and synchrotron based work.
https://research.pasteur.fr/en/emember/patrice-vachette/

Frank Gabel (Institut de Biologie Structurale, Genoble) – currently the leading expert in SANS/contrast variation and SANS/SAXS/NMR hybrid structural work.

Biochemistry Bldg, G08
Faculty of Science
The University of Sydney
NSW 2006 Australia

T +61 (0)4 3275 3933
E jill.trewhella@sydney.edu.au
http://sydney.edu.au/science/people/jill.trewhella.php

ABN 15 211 513 464
CRICOS 00026A

http://www.ibs.fr/research/research-groups/extremophiles-and-large-molecular-assemblies-group/small-angle-scattering-649/

Lois Pollack (Cornell) – leader in RNA and RNA-protein studies, time-resolved SAXS who was also highly recommended by Dmitri Svergun.
http://www.aep.cornell.edu/people/profile.cfm?netid=lp26

Dina Schneidman (Hebrew University) – currently publishing work with a rigor and completeness that it provides a model for the committee who was also highly recommended by Andrej Sali. http://www.huji.ac.il/dataj/controller/ihoker/MOP-STAFF_LINK?sno=70741876&Save_t=

I appreciate the opportunity to work with this important committee to continue to support high quality SAS and its utilisation in biomolecular structural analysis. Widely agreed and utilized data and model validation criteria are essential for SAS to make its appropriate contribution to the growing field of integrative/hybrid structural biology.

Yours sincerely,

Jill Trewhella, PhD

cc SASvtf members

WORLDWIDE PDB
PROTEIN DATA BANK
wwpdb.org • info@wwpdb.org

February 18th 2017

Jill Trewhella, Ph.D.
Chair, wwPDB SAS VTF

Gaetano Montelione, Ph.D. and Micheal Nilges, Ph.D.
Co-Chairs, wwPDB NMR VTF

E-mails: jill.trewhella@sydney.edu.au, guy@cabm.rutgers.edu, michael.nilges@pasteur.fr

Dear Jill, Guy, and Michael:

We write to request formally that you three and select members of the wwPDB Small-Angle Scattering Validation Task Force and the wwPDB NMR Validation Task Force meet as soon as is practicable to review the state-of-the-art and develop and promulgate recommendations regarding validation of integrative/hybrid methods NMR-SAS refined atomic coordinate structures. Your guidance is urgently required to ensure that integrative/hybrid NMR-SAS structures entering the Protein Data Bank archive are subject to rigorous community agreed validation procedures.

We are most grateful for your efforts on behalf of the Protein Data Bank archive and the wwPDB partnership, and stand ready to provide any help that we can in terms of meeting logistics, etc.

Yours faithfully,

Stephen K. Burley, D.Phil., M.D.
Distinguished Professor
Rutgers, The State University of New Jersey
Director, RCSB Protein Data Bank

Savmeer Velankar, Ph.D.
Team Leader, EMBL-European
Bioinformatics Institute
Head, PDBe

John L. Markley, Ph.D.
Steenbock Professor of Biomolecular Structure
University of Wisconsin-Madison
Head, BMRB

Haruki Nakamura, Ph.D.
Director, Institute for Protein Research,
Osaka University
Head, PDBj

WORLDWIDE PDB
PROTEIN DATA BANK
wwpdb.org • info@wwpdb.org

February 18th 2017

Jill Trewhella, Ph.D.
Chair, wwPDB SAS VTF

E-mail: jill.trewhella@sydney.edu.au

Dear Jill:

We write to request formally that the wwPDB Small-Angle Scattering Validation Task Force meet as soon as is practicable to review the state-of-the-art and develop and promulgate recommendations regarding validation of data and metadata coming from small-angle scattering (SAS) studies of biological macromolecules.

Your guidance is urgently required to ensure that SAS data contributing to determination of integrative/hybrid NMR-SAS structures entering the Protein Data Bank archive are subject to rigorous community agreed validation procedures.

We are most grateful for your efforts on behalf of the Protein Data Bank archive and the wwPDB partnership, and stand ready to provide any help that we can in terms of meeting logistics, etc.

Yours faithfully,

Stephen K. Burley, D.Phil., M.D.
Distinguished Professor
Rutgers, The State University of New Jersey
Director, RCSB Protein Data Bank

Savmeer Velankar, Ph.D.
Team Leader, EMBL-European
Bioinformatics Institute
Head, PDBe

John L. Markley, Ph.D.
Steenbock Professor of Biomolecular
Structure
University of Wisconsin-Madison
Head, BMRB

Haruki Nakamura, Ph.D.
Director, Institute for Protein Research,
Osaka University
Head, PDBj

Jill Trewhella
Professor Emeritus

19 February 2017

Dina Schneidman-Duhovny,
The Hebrew University of Jerusalem
Department of Biological Chemistry
Jerusalem, Israel

Email: duhovka@gmail.com

Dear Dina,

I would like to invite you to join the Small Angle Scattering Validation Task Force (SASvtf). This committee was originally formed to make recommendations on world-wide Protein Data Bank (wwPDB) acceptance of models based on SAS studies, and if so, what types of data and validation should be included. With the growing numbers of integrative or hybrid models (IHM) and the formation of the IHMvtf, the work of the SASvtf has become even more critical.

Initial reports from the SASvtf and IHMvtf are attached that provide more details of the work to date from these committees. At this time, an updated set of recommendations for publishing the results of biomolecular SAS experiments, including data and model validation, is being prepared to be reviewed by the SASvtf and discussed at a virtual meeting to be scheduled early in May, 2017.

I hope that you will be able to serve on this important committee. With your positive response, you will be contacted regarding the scheduling of the May meeting.

Yours sincerely,

Jill Trewhella, PhD

cc: wwPDB Leadership Team
SASvtf Members

Biochemistry Bldg, G08
Faculty of Science
The University of Sydney
NSW 2006 Australia

T +61 (0)4 3275 3933
E jill.trewhella@sydney.edu.au
http://sydney.edu.au/science/people/jill.trewhella.php

ABN 15 211 513 464
CRICOS 00026A

Letter to SASvtf, sent to all participants 2/28/17, 3:53pm EST

Colleagues

I am pleased to welcome the new members to our Small-Angle Scattering validation task force (SASvf, complete list of members appended below) and look forward to working with the expanded group on the next steps towards establishing a set of SAS data and model validation criteria for biomolecular structural studies that can have the broad support of our community and ensure that SAS can make its proper contribution to the growing field of integrative/integrative structural biology for which our sister wwPDB task force (the IHMvtf) produced a first report in 2015 (Sali et al. "Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop," *Structure 23*, 1156-1167, 2015) and which is proceding toward implementation of the recommendations made therein.

For the benefit of the new members of the SASvtf, and perhaps a handy reminder for others, I restate here information from my letter to the task force of 3 February, 2017 with some updates regarding where things are with respect to preparing for our May virtual meeting, for which Maggie Gabanyi will be contacting you regarding scheduling.

The May meeting has the following agenda:

1. review progress to date on the implementation of our recommendations as outlined in our 2013 report in *Structure* and plan next steps; and
2. consider in detail the fourth recommendation from the SASvtf; that is "criteria need to be agreed upon for the assessment of the quality of deposited data and the accuracy of SAS-derived models, and the extent to which a given model fits the SAS data."

With respect to item 2, in our 2013 report (Trewhella et al, "Report of the wwPDB Small-Angle Scattering Task Force: Data Requirements for Biomolecular Modeling and the PDB," *Structure 21*, 875-881, 2013) we expanded upon preliminary guidelines (Jacques et al., "Publication Guidelines for Structural Modelling of Small-Angle Scattering Data from Biomolecules in Solution," *Acta Cryst. D68*, 620-626, 2012) that addressed **sample quality**, **data acquisition and reduction**, **presentation of scattering data and validation**, and **modelling** for biomolecular SAS. These preliminary guidelines were developed in consultation with the IUCr SAS Commission and other experts in the field and formally adopted by the IUCr Journals. They have subsequently been used by a growing cohort of SAS researchers.

I am in the process of drafting an update of the preliminary publication guidelines to a v1.0 that would incorporate the additional points made in from our 2013 report, as well as input from the IUCr SAS Commission and other experts in the field. I will circulate this draft to the SASvtf in the next few weeks for comment. At our virtual meeting we would review all input and agree a final set of guidelines, including specific, implementable recommendations for both data and model validation, which we would formally recommend as the SASvtf.

I plan to present the revised, updated publication guidelines at the IUCr Congress in Hyderabad in August during an open meeting of the IUCr SAS Commission that I have called for as Chair of the SAS Commission. Subsequently we will submit the paper for publication describing the revised guidelines and the basis for recommended updates.

Kind regards, jt

Letter to SASvtf, sent to all participants 2/28/17, 3:53pm EST

SASvtf members

Andrej Sali	sali@salilab.org
Dina Scheidman	dina.schneidman@mail.huji.ac.il
Dmitri Svergun	svergun@embl-hamburg.de
Frank Gabel	frank.gabel@ibs.fr
Jill Trewhella (Chair)	jill.trewhella@sydney.edu.au
John Tainer	jtainer@mdanderson.org
Lois Pollack	lp26@cornell.edu
Masaaki Sugiyama	sugiyama@rri.kyoto-u.ac.jp
Patrice Vachette	patrice.vachette@i2bc.paris-saclay.fr
Torsten Schwede	torsten.schwede@unibas.ch
Wayne Hendrickson	wayne@convex.hhmi.columbia.edu

PROFESSOR EMERITUS JILL TREWHELLA

THE UNIVERSITY OF SYDNEY

Faculty of Science | The University of Sydney | NSW | 2006
E jill.trewhella@sydney.edu.au | W http://sydney.edu.au

CRICOS 00026A