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

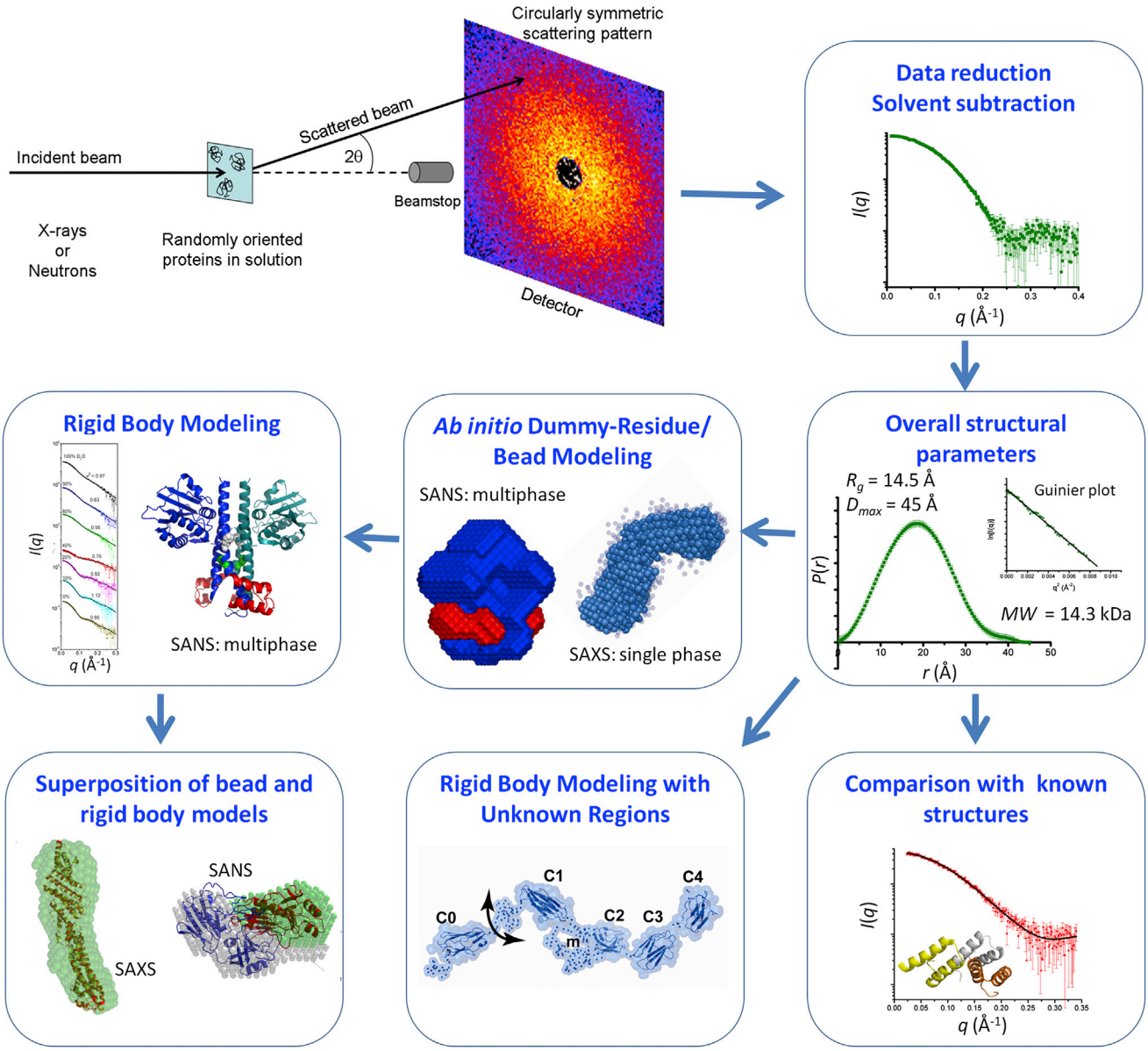

Figure 1. Roadmap of SAS Data Collection and Analysis

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.