Analysis of biostructural changes, dynamics, and interactions - Small-angle X-ray scattering to the rescue
Vestergaard, Bente

Published in:
Archives of Biochemistry and Biophysics

DOI:
10.1016/j.abb.2016.02.029

Publication date:
2016

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):

Download date: 19. sep., 2019
Analysis of biostructural changes, dynamics, and interactions — Small-angle X-ray scattering to the rescue

Bente Vestergaard

Department of Drug Design and Pharmacology, University of Copenhagen, Jagtvej 162, 2100 Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 18 January 2016
Received in revised form 17 February 2016
Accepted 26 February 2016
Available online 2 March 2016

Keywords:
Small angle X-ray scattering
Complex structural analysis
Intrinsically disordered proteins
Advanced sample environments
Hybrid biostructural analysis

ABSTRACT

Solution small angle X-ray scattering from biological macromolecules (BioSAXS) plays an increasingly important role in biostructural research. The analysis of complex protein mixtures, dynamic equilibria, intrinsic disorder and evolving structural processes is facilitated by SAXS data, either in stand-alone applications, or with SAXS taking a prominent role in hybrid biostructural analysis. This is not the least due to the significant advances in both hardware and software that have taken place in particular at the large-scale facilities. Here, recent developments and the future potential of BioSAXS are reviewed, exemplified by numerous examples of elegant applications to challenging systems.

© 2016 The Author. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Frontiers in biostructural analysis

Prime challenges in biostructural analysis today include the investigation of structural changes, dynamics and polydispersity. The accurate and detailed description of biomacromolecules, derived from high-resolution structures and primarily originating from macromolecular crystallography (MX), has provided the research community with a wealth of insight, greatly enabling our current understanding of biomolecular function. Biomacromolecules, however, are inherently dynamic at several time- and length-scales [1], and this dynamic behavior is crucial for their biological function. The cell is a crowded and ever-changing environment where proteins, lipids, nucleic acids, polysaccharides and other biomolecules interact in a structurally responsive and adaptive manner. A single protein structure should thus not be considered as a three-dimensional rigidly defined entity, but rather be understood as a spatiotemporal distribution of an ill-defined number of conformational states, and this ensemble of conformations defines the biological function of the protein. Likewise, when considering macromolecular interactions (e.g. protein:protein interactions [2]), structural polydispersity plays a significant role: often complexes are only partially and/or transiently formed, and complex formation in addition may induce or require different levels of conformational changes of the individual protomers. One may say that the functionality of macromolecular interactions is defined by a highly refined interplay between two macromolecular entities, each defined as a complex ensemble of structures, and that this interplay introduces an additional level of structural complexity, not exhibited by each individual structure prior to the encounter.

This view on structures challenges biostructural analysis in general, and high-resolution structural investigation in particular. Small angle scattering (SAS), on the other hand, is uniquely suited for these experimental endeavors. In a SAS experiment the sample is in the liquid state, and the experiment can be performed under physiologically or otherwise experimentally relevant conditions, since there are no particular requirements to sample preparation. SAS is a low-resolution method, and hence by no means replaces, but rather uniquely complements high-resolution biostuctural analysis. As a consequence, the SAS community is experiencing an almost explosive development in numerous ways: the available software and hardware rapidly advances, the complexity of the scientific questions that are addressed is continuously increasing [3], and there is an ever-

Abbreviations: SAS, small angle (X-ray/neutron) solution scattering; SAXS, small angle X-ray scattering; SANS, small angle neutron scattering; BioSAXS, biological solution SAXS; BioSANS, biological solution SANS; SEC-SAXS, size exclusion chromatography coupled with SAXS data collection; TR, time resolved; IspyB, Information System for Protein Crystallography Beamlines; IspyBR, IspyB for BioSAXS; MW, molecular weight; SASDB, Small Angle Scattering Biological Data Bank; PDB, protein data bank; RF2, polypeptide release factor 2; IDP, intrinsically disordered protein; NMR, nuclear magnetic resonance; MX, macromolecular crystallography; WAXS, wide angle X-ray scattering.

* This article is part of a Special Issue entitled Protein Crystallography, edited by Ana Camara-Artigas and Jose Antonio Gavira. E-mail address: bente.vestergaard@sund.ku.dk.
2. Increasing data quality and decreasing sample quantity

The intensities of X-rays and, to a lesser extent, neutrons at international large-scale facilities are dramatically increasing. This drives impressive hardware advances at the leading established and upcoming facilities within both biological solution small angle X-ray scattering (BioSAXS) [5–11] and neutron scattering (BioSANS) [12] facilities. Not only have the beam parameters (such as brilliance and positional stability) constantly improved, so have the front-end hardware and software developments, the latter dramatically increasing user-friendliness. At several BioSAXS facilities robotics ensure low sample consumption and robust, rapid sample mounting [7,13,14]. Thereby, not only is sample/data-throughput significantly improved, also, the standardized mounting minimizes user mistakes and includes an optimized cleaning procedure, while oscillation of the sample in the capillary diminishes radiation damage, with the net effect of improved data quality. The sensitivity and speed of detectors has also undergone revolutionary development, and one consequence is that now, multiple data-frames from very short exposures are collected for the data collection from a single sample, rather than performing one long exposure. This enables comparison of individual data-frames, and hence detection of potential radiation damage from the high-brilliance X-ray beams, thereby ensuring that the final dataset for a given sample is averaged only from the frames devoid of radiation damage.

A highly useful tool which was recently developed, and which, among many other useful applications, can detect the potential onset of radiation damage during multiple exposures, is the correlation map (CorMap) [15]. With this tool, it is possible to perform pair-wise or multiple comparisons of data-curves, independent of the error estimates. The method is highly sensitive, and has a large range of important applications, some of which are mentioned in this review. One example, applied during data collection, is the continuous monitoring of the cleanliness of the sample cell. If collecting a large number of datasets, repeated buffer measurements can be compared, and should correlate perfectly, if the sample cell remains clean. The implementation of such automated procedures at the advanced beamlines greatly enhances the output for users visiting the facilities.

Also, the greatest part of basic data reduction, processing and analysis has been automated. These routines have been included in a major pipeline of data evaluation which calls individual advanced software packages, providing the user with on-the-spot information about data quality, biophysical parameters and ab initio structures [16]. These and other automated procedures are also included in IspyBB (Information System for Protein crystallography Beamlines (IspyB) for BioSAXS) [17]. IspyBB integrates data-management from the point of preparing the samples (strategy for data collection, calculation of the needed sample quantities etc.), over logging and controlling the data collection to providing the results of the initial automated data analysis in a GUI. In addition IspyBB enables access to and comparison with relevant data from the same or previous data collections [17]. With low sample consumption and rapid data collection BioSAXS users collect tens, hundreds, and sometimes thousands of datasets within a project period, which increases demands for comparative data evaluation, or even data archiving and tracking. It is planned to expand the program, such that the measured data guides additional sample preparation using liquid-handling robotics at high-throughput crystallization facilities. As an example, if the SAXS data reveal that certain conditions promote complex formation and minimize unspecific aggregation, the following data collection will use these or further optimized experimental conditions [17] which ultimately enables a scanning of a multidimensional experimental space, searching for relevant structural states of the investigated macromolecules [18].

Even lower sample consumption can be achieved by the use of microfluidic sample environments [6,18–21]. Some microfluidic systems aim at providing versatile off-the-shelf sample environments for standard screening purposes, however, microfluidics have also in several cases been applied in a customized setup with a particular purpose, such as the screening of structural changes in lipidic mesophases induced by the experimental conditions applied during membrane protein crystallization attempts [22], or the on-chip dialysis setup enabling in situ sample concentration and buffer exchange for fragile protein samples [23]. A particularly useful application when combining microfluidics and SAXS is in time-resolved (TR) studies. In fact, microfluidic mixing, both stopped-flow and continuous flow, has been used for more than two decades for the TR study of protein and nucleic acid un- and refolding [24–27]. TR studies are applied today with as low as nl sample consumption [20] and to increasingly complex samples, such as the intermediate nucleosome states during DNA unwinding [28] or intermediate filament formation formed under ionic gradients [29].

3. In-line sample purification and orthogonal data

More slowly evolving mixtures can be efficiently analyzed using SEC-SAXS (size exclusion chromatography coupled with BioSAXS data collection, Fig. 1) [5,30,31]. That is, SAXS data are collected as the sample is purified, by directing the liquid flow directly from the purification column and through the X-ray beam. As simple as this idea may sound, this development is by no means trivial, and has been made possible in part because the increasing intensities at synchrotron beamlines allow for data collection in very short time-intervals. That is, a very large number of data-frames are collected over the elution profile, each frame of a sufficient quality to allow for individual data analysis. Individual data-frames are hence comparatively evaluated, and data-frames from a single eluting species can be isolated in an (semi-) automated manner [5,32–34]. SEC-SAXS can dramatically improve data quality from aggregation prone proteins. If small amounts of un-specific aggregates co-exist with the molecule of interest, this compromises the SAXS data, and in many cases makes any attempt of further analysis futile. With SEC-SAXS, however, the sample of interest is separated from these aggregates and data are collected before the purified sample re-aggregates. SEC-SAXS is also very useful in the case of partial complex formation, where only a fraction of protomers form complexes, and these complexes co-exist with the non-complexed individual proteins (Fig. 1).

Analysis of such mixed states can be very difficult without significant prior information, but SEC-SAXS often allows sufficient separation of the species such that spectra can be obtained from the pure species. Recent examples of such successful analysis are e.g. the studies of prion protein in complex with antibodies [36], co-eluting monomers and dimers of intrinsically disordered alpha- synuclein [37], monomeric, oligomeric or degradation products of fibrinogen [39], MnME:MnmG complexes of different stoichiometric states [38] or co-existing apo- and ligand-bound forms of Arabidopsis thaliana acyl acid-amido synthetase [31]. A particularly useful application is in the analysis of membrane proteins [39]. Detergent solubilized membrane protein samples are a mixture of the protein:detergent complexes, soluble detergent and varying amounts of detergent micelles. Since successful analysis of SAXS data crucially relies on correct background subtraction, SAXS data...
The UV-trace from the eluting protein and the correlated SAXS data at specific time-points of elution. The UV-measurement provides the sample concentration, which can be used as input to the SAXS data evaluation, since the scattering intensity is proportional to sample concentration. In one setup right-angle light scattering and refractive index detectors are also combined ex situ (but in an integrated pipeline) from the SAXS setup [5] thereby enabling independent estimates of the molecular weight (MW) and hence oligomeric state of the eluting protein species. Another in situ cell successfully combines SAXS, UV, Raman and fluorescence measurements [40] and similar and even more advanced environments are in development. Data collection from orthogonal data sources is a sound principle of any structural or biophysical evaluation, and for BioSAXS data in particular. As outlined below, the analysis of the radially averaged 1-dimensional SAXS data curve is burdened by ambiguity [41], which is greatly reduced when incorporating complementary information.

4. Basic analysis and ab initio modeling: from 1D to 3D without getting lost

To outsiders and newcomers, the process of analysis from the radially averaged one-dimensional scattering curve (scattering intensities versus the momentum transfer range $q$; $q = 4\pi \sin(\theta)/\lambda$; $2\pi$ is the scattering angle and $\lambda$ is the X-ray wavelength) to descriptions of mixture compositions, structural dimensions, three-dimensional models, structural descriptions of disordered proteins, structural conversions, etc. can seem like magic. And it must be emphasized that any modeling based on SAS curves is associated with ambiguity. There will be several models that equally well will describe a given scattering curve. This fact, however, by no means excludes the possibility to perform even detailed modeling based on scattering data, but leaves the SAS-user with the responsibility to very carefully report how the SAS data were analyzed [42], and to evaluate the suggested SAS-based models by alternative means. Also in this light, an important progress in the community, which eases cross-validation, elaboration and collaboration based on BioSAXS data, is the establishment of a curated repository, the Small Angle Scattering Biological Data Bank (SASBDB). SASBDB enables free access to scattering data, experimental information and derived models [43], following the guidelines recommended by the SAS task force established under the protein data bank (PDB) [44].

A key point for extending SAS data evaluation to advanced levels is the inclusion of prior knowledge, when analyzing data. Prior knowledge in this context can be many things. One example is the knowledge that a protein structure at low resolution in solution can be described as an undisrupted particle with a near-uniform electron density. Although such a fact may come across as a simple and very basic fact, such restraints, when performing modeling, significantly reduce the ambiguity of modeling and guide programs to robust solutions. Evidently, when in possession of further knowledge such as the particle concentration, the protein sequence, the oligomeric state or even partial high-resolution structures or homology models (e.g. of single domains in a multi-domain protein), such knowledge can be efficiently incorporated into data analysis routines. And, as further discussed below, the inclusion of complementary data from orthogonal methods greatly empowers SAS data analysis.

To obtain information from a sample the protein samples must be carefully matched by background samples, that is, samples of the exact same composition as the protein sample, only devoid of protein. These samples are measured and used for background subtraction. However trivial this may sound, incorrect background subtraction can often be the key reason for futile or (even worse)
in-correct data analysis, and hence is very explicitly mentioned here. There can be several reasons for obtaining inadequate data for background subtraction. One example was already mentioned above in context of membrane protein analysis, where the composition of soluble micelles is influenced by the presence of the membrane protein, thereby complicating the measurement of the background sample. Another reason can be (sometimes rather extensive) interactions between buffer components and the protein (or other macromolecule), and hence it is recommended that sample dialyses (rather than the originally produced buffers) be used for background measurements.

While highly complex analysis can be performed from SAS data, it is crucial to provide significant attention to all basic, initial data evaluation (including evaluation of the background subtraction) [42]. Although, as mentioned above, several beamlines now provide automated procedures for on-the-fly data evaluation, this by no means diminishes the need for a subsequent thorough manual evaluation of the data. Rather, the automated procedures are highly valuable because they provide the user with an overview of both sample quality and structural parameters while collecting data, thereby facilitating optimal use of precious sample and valuable beamtime.

After background subtraction, it should be assessed which part of the curve contains useful information. Even though data may have been collected to relatively high momentum transfer range, the data may not contain information if the noise levels are high and the uniqueness of the curve is low, which can be evaluated [45], and such an evaluation is implemented in some, but not all beamlines.

At the initial points of analysis it is possible to obtain a large number of basic biophysical parameters for the particle of interest (such as the particle MW and hence oligomeric state, radius of gyration (Rg), maximal dimension (Dmax) and overall shape, the latter originating from the indirect Fourier transform of the data, yielding the pair distance distribution function P(r)). It is always recommended to collect data from several sample concentrations, since comparison of such data will evidence the presence of potential interactions (repulsion or attraction) between the molecules or concentration-dependent oligomerisation of the particles. At this point of the analysis it is often possible to detect if the sample quality is inadequate for further analysis (e.g. presence of aggregates, impurities, degradation and similar problems). In fact, SAXS data brutally exposes the sample quality with very high sensitivity, which is one of many reasons for the growing interest in the bio-pharmaceutical and biotechnological industry to use BioSAXS as an advanced screening tool in formulation development [46]. Crucial factors in formulation development include the assessment of interactions, structural and/or colloidal stability and early detection of aggregates formation, and this can be evaluated by inspection of the abovementioned basic biophysical parameters. With the described development in sample handling robotics BioSAXS data collection can be applied in a high-throughput mode, and comparative basic data evaluation will highlight formulation differences in a highly sensitive manner.

For a structural biology laboratory, further details are pursued from continued analysis of the data. Very often low-resolution ab initio modeling is pursued. Probably, the all-dominant method for such modeling is the bead-modeling approach advocated by Svergun and co-workers [47], DAMMIN [48] and DAMMIF [49]. These programs apply a simulated annealing protocol to generate shapes (ensembles of beads) with protein-like properties (in terms of connectivity, electron density) that will fit the scattering curves. To reduce the challenge of ambiguity multiple constructions of models and subsequent comparison of the spatial discrepancies among models and averaging yield the final models [50]. A recent approach addresses the ambiguity problem based on the scattering curve, i.e. prior to modeling attempts. Based on a large number of constructed shapes a landscape of computed scattering curves is used to identify the number of neighboring (i.e. similar) curves for a given experimental scattering curve [51]. Curves with a low number of neighbors can hence be said to be associated with a reduced level of ambiguity when using ab initio modeling.

Ab initio modeling can principally only be performed on data from monodisperse samples, since the program will search for one model to describe the scattering curve, which hence likewise must derive from one scattering particle. A more recent development, however, allows for the assessment of shapes from equilibrium systems of monomers and symmetric homo-oligomers [52]. The analysis of data from more complex mixtures demands different approaches as discussed in the paragraph below.

5. Working with prior structural information

In the context of this special issue, the inclusion into SAXS analysis of prior structural information from MX experiments is a very relevant topic (Fig. 2). While MX captures atomic resolution details, it is methodologically impossible to capture the dynamics of the solution state including e.g. the large conformational space covered by multi-domain proteins with extensive linkers [56,57] or mixtures of transient and/or flexible complexes [58,59]. In these cases BioSAXS data have the capacity to uniquely complement a high-resolution structure. As such, BioSAXS is extensively used to (in-) validate and elaborate on the understanding of particular conformations of multi-domain proteins or protein complexes, observed in crystallo, e.g. the evaluation of the ubiquitin binding site on proliferating cell nuclear antigen [60] or the orientation of individual subdomains and flexible loop areas in the multi-domain complement component C4b [61]. For such analyses, the central tool is a program that calculates the theoretical solution scattering pattern based on an available high-resolution structure. Several such programs exist. A very popular program is CRYSOL, developed by Svergun et al. [62]. The program was the first to include a general description of the solvent layer organized near the surface of the macromolecule, which must also be included in the evaluation of the total particle scattering. This is a non-trivial question, which originally was addressed in a combined neutron and X-ray scattering study data [63] and which has been recently re-addressed [64]. Numerous other applications exist for calculating the theoretical scattering curves. There is some variation in these programs concerning the calculation of the scattering pattern, but more extensively debated is the principle behind the description of the solvent layer. While CRYSOL uses a multipole expansion and spherical harmonics to calculate a uniform solvent layer with an extension and an average electron density that are both refined in the process [62], the program AXES includes explicit solvent modeling, and in addition allows for the fitting to numerous explicit input structures (both the individual structures and the average scattering pattern from all structures) [65]. The AquaSAXS server also enables the incorporation of several pdb-files in the fit, and a choice between the use of the AquaSAXS solvent model (which is defined by orientable dipoles) [66] or a solvent layer defined by the principles behind the FoXS approach. The FoXS program bases the solvent description on a calculation of the atomic solvent accessible area [67] and a novel extension of the program (MultiFoXS) includes a calculation of a large number of potential conformations that are accessible to a flexible protein and selects ensembles based on the branch-and-bound method [36]. The program Bayesian Ensemble SAXS (BE-SAXS) describes the solvent layer following the same principles as in FoXS (personal communication), but this program is developed for ensemble
modeling rather than evaluation of high-resolution structures (see below). And finally, the new program WAXS implements an explicit solvent model, calculated by molecular dynamics simulations [68,69].

While the evaluation of experimental SAXS data can be aided by the inclusion of such prior structural information (and/or the prior structural information can be evaluated by the SAXS data), other programs, developed for protein structure prediction, include SAXS data (or other data sources) as restraints in their protein structure modeling. One such program, called PHAISTOS [70], has a SAXS module using a coarse-grained representation of the protein (each amino acid represented by two dummy atoms) and the Debye formula to calculate the scattering patterns [71] reducing computational time by GPU parallel threads, but without the inclusion of a model for the solvent layer [70]. Otherwise following approximately the same principles, the protein structure prediction suite BCL implements an explicit solvent model in their BCL::SAXS module [72].

The correct theoretical representation of the scattering from a given high-resolution model is prerequisite for any further modeling including this prior information. As mentioned, controversial crystal structures have been evaluated by solution scattering data (see examples above) but it is also possible to extend available partial structures by SAS-based modeling. If e.g. the monomeric structure has been crystallized, and SAS data exist for the homo-dimer, docking algorithms, attempting to identify the correct protein:protein interaction sites increase their performance when guided by SAXS data [73–76]. Likewise if one conformation of a given protein has been crystallized (e.g. the ligand-bound form), the conformational changes in solution, e.g. for the apo-form, can be evaluated using rigid body modeling approaches. In rigid body modeling rotation of individual domains and/or protomers of a complex [77] are guided by e.g. molecular dynamics simulations [78]. Rigid body modeling can also include contrast variation data, e.g. applicable to complexes with DNA/RNA, or where individual protomers of a complex are deuterated [79]. In a recent variant of rigid body refinement, normal mode analysis defines ‘pseudo-domains’ in a given crystal structure, followed by a hierarchical refinement of large to local-scale movements, fitting to the SAXS curve [80].

If a given crystal structure is lacking electron density for parts of the structure (e.g. loops or termini), these can be modeled based on SAS data [77], and also more extensive missing regions can be modeled in a hybrid high-resolution/bead-model/explicit model approach [81]. A particularly useful application of this approach is in the modeling of membrane protein structures, embedded into nanodisks [82]. There, the availability of the complementary information from neutrons and X-rays is prerequisite for the model development. This is one example of how orthogonal data can reduce ambiguity, and hence increase the level of complexity that
Fig. 3. Intrinsic disorder is captured in BioSAXS analysis. Ensembles of structures, representative for the distribution of the very large number of conformations present in solution, can be derived from SAXS data, here represented in a pioneering study of tau protein [105]. Entry ID PED7AAD.

can be addressed from a given experiment.

And back to basics: for any modeling, the correct assessment of basic data characteristics, such as perfect background subtraction, evaluation of the mono- or polydispersity of a given sample [42], assessment of the meaningful data-range [45], and optimal evaluation of discrepancies between models and data [15] remain crucial. However sophisticated a given modeling approach may be one can only derive the information that actually resides in the data. And irrespective of the quality of a given model, one must never forget the inherent ambiguity in SAS modeling, leaving the responsibility of independent model evaluation by complementary experiments and/or evaluation based on prior knowledge on the researcher.

6. Polydisperse samples

As mentioned above, in most tools for SAXS-based 3D modeling it is assumed that data are from monodisperse samples. However, polydispersity is an inherent structural parameter of many macromolecules, such as e.g. intrinsically disordered proteins (see below), multi-domain proteins, partially formed complexes and developing systems (Fig. 4).

An important extension of the rigid body modeling tools is found in SASREFMX [83] which allows modeling against data from polydisperse samples, e.g. transient or weak complexes, where the protomers exist in a mixture of bound and free states. To lower the ambiguity level, data recorded at different conditions (e.g. varying protein concentration) should be included. If the complexity of the sample extends further, e.g. when analyzing a partial complex formation which induces structural changes of individual protomers [58], or when analyzing evolving mixtures such as e.g. protein fibrillation reactions [84–89], it is necessary to apply different tools. A SAXS spectrum is the product of the structure factor and form factor contributions. While the form factor describes the scattering contribution from the individual scatterer (e.g. the protein) the structure factor describes the interaction between particles (i.e. repulsive or attractive effects, causing an ordering of the particles (proteins) in solution). If the structure factors are negligible, SAXS data are additive. This means that in these ideal cases, data from a mixture of different proteins will equal the sum of the scattering curves from the individual scattering components, weighted by their relative volume fractions. In principle, it is hence possible to decompose such mixture data into the pure spectra from individual species, and indeed there are tools to do this. The program OLIGOMER can calculate the volume fractions (relative weights) of individual components in a mixture, if the (theoretical or experimental) SAXS profiles for each component are available [90]. When in possession of data where the relative concentrations of the individual species are varied, e.g. by varying the protein concentration, temperature or other experimental conditions, or by measuring along a given reaction coordinate over time or via titration, it is also principally possible to derive the pure spectra of individual species, which are not known prior to the experiment (e.g. Refs. [84–89]). As above, either prior knowledge about a number of components must be available, and/or their corresponding volume fractions. Prior knowledge about the pure spectra can be either experimental or theoretical (from e.g. homology models or docking models), and prior knowledge about volume fractions can be based on spectroscopy or other orthogonal data sources. Indeed the decomposition is burdened by ambiguity thus it is important to use orthogonal data to confirm the validity of the procedure. Such complex data can also with advantage be analyzed using chemometrics based approaches for the actual decomposition procedure (see e.g. Refs. [59,91]). Once decomposed, the pure spectra can evidently be analyzed and used for modeling, as described above and hence the actual decomposition is the bottleneck for analyzing data from complex mixtures.

Fig. 4. Dynamic protein structural equilibriums. Structural characterization of heterogeneous macromolecular states is challenging. SAXS is a versatile method, applicable to the analysis of macromolecular systems that cannot easily be characterized by other structural biology methods [128]. The hypothetical monomeric protein (green) has an intrinsically disordered N-terminus (lighter green, a number of potential conformations are shown). SAXS is very useful for low-resolution characterization of the solution structure [129,130], for evaluation of existing high-resolution structures [131] and for modeling of intrinsic disorder [109]. The N-terminus becomes ordered upon dimer formation (red-purple). The monomer-dimer equilibrium and protein structural changes can with advantage be characterized by SAXS analysis [128]. The protein fibrillates under particular experimental conditions (fibrils are sketched in orange colors (not to scale with the monomeric protein)). These experimental conditions can be applied during SAXS analysis of the fibrillation process [132], since there are very few restrictions on SAXS sample preparations.
7. Protein fibrillation, analyzed by SAXS and SANS

The formation of fibrils from amyloid proteins associates with at least 20, often fatal, diseases, including the neurodegenerative Alzheimers and Parkinsons diseases [92,93]. A fibrillating amyloid protein is a rather extreme example of a polydisperse, developing system, which is inherently difficult to analyze structurally (Fig. 4). The protein evolves from its native state into \( \mu \)-long extended structures, via intermediate structural states. Such intermediate states are suggested to associate with cytotoxicity [94], which only exist in equilibrium with the native and fibril states, and hence cannot be purified from the reaction mixture without the risk of perturbing the structure. It is thus a prerequisite for analysis that it is possible to work on undisturbed mixtures in solution, which applies to SAXS and SANS. In addition, the reaction is highly sensitive to the experimental conditions applied [95,96], which makes it a significant advantage that there are very few restrictions on the experimental conditions that can be applied in a SAXS analysis. Given the very broad range of resolution, that is covered in SAXS data, the method is particularly suitable for fibrillation analysis, and SAXS data can bridge between high-resolution and low-resolution data, such as we demonstrated in a study of the hierarchically organized fibrils of a heptapeptide fragment of prion protein [97]. In a pioneering SANS study on the development of \( \beta \)-lactoglobulin fibrillation, the data were fitted to a model with charged spheres and long cylinders (representing monomers and fibrils respectively), yielding the concentration profiles and hence a model of the fibrillation kinetics [98]. Our first SAXS-based fibrillation study of human insulin provided the first-ever solution structure of a transient oligomeric amyloidogenic structure, determined without perturbing the reaction mixture [88]. We used data decomposition to isolate the scattering contribution from the intermediate structure, which was never present in solution alone, and we applied \textit{ab initio} modeling to both the intermediate and fibril structures. The presence of such an intermediate species was also observed for glucon fibrillation [99] and later for \( \alpha \) and later for DAT [100]. In the latter case, the observed dimensions of intermediate and fibril structures made us suggest that fibrils would form by oligomer stacking, a model which was corroborated in a later study [87], where the intermediate structures were stabilized by a small-molecule compound [101]. One great challenge in investigating fibrillation processes is the fact that several pathways of aggregation may co-exist and, that small experimental changes will induce also structural/pathway changes thus several oligomeric forms may form. Indeed, structurally different \textit{aSN} oligomeric forms have been observed by SAXS by others than us [102], and as of now, it remains elusive which of such intermediates, if any, are relevant for establishing a molecular understanding of cytotoxic effect [94]. In other cases, our SAXS analyses have shown that the fibrillation proceeds without accumulation of significant amounts of intermediate oligomers [85,97]. Rather, in the case of transthyretin fibrillation, we revealed that a highly unfolded monomer co-exists and assumingly interchanges with the amyloid protofibrils [85]. If indeed such a fibrillar state interchanges with the soluble state, this may be a key for understanding the catalytic effects observed when feeding with fibril material, and/or in the kinetics profiles [103].

8. Intrinsically disordered and highly flexible proteins

Two classes of proteins escaping structural elucidation by MX are multi-domain proteins with extensive flexible linkers, and intrinsically disordered proteins (IDPs). In the former case functionality closely links with the extensive flexibility and the 3D space covered by the flexible systems, and this phenomenon is not described by single crystal structures (although evidently the high-resolution information about the individual domains is still highly valuable) [56]. In the latter case, the IDP proteins lack structural characteristics that are well described by classical structural terms. The range of intrinsic disorder can vary from termini or loops to larger fractions or entire proteins. This, however, does not imply that these structures are random. Rather, these proteins should be described as an ensemble of a very large, ill-defined number of conformations, each occurring with certain probability. This probability distribution is evidently as defined by amino acid sequence as any type of structure, however no individual conformation describes the structure:function relationship of this type of protein, but the weighted sum of conformations does.

SAXS is well suited for the investigation of this type of molecules. The Kratky plot (\( q^2I(q) \) versus \( q \)), can be used to visualize the (lack of) compactness of a protein (see e.g. Ref. [104]). A folded protein will display an approximately \( 1/q^2 \) decreasing scattering intensity at higher \( q \)-range, while a random chain will display \( 1/q^2 \) behavior. This means that these two extremes are very easily distinguished in a Kratky plot. For in-between cases, such as a partially folded protein or an intrinsically disordered protein, the Kratky plot will display distinct features that can easily, qualitatively be recognized and one can readily visualize the transition of e.g. a structured protein to a chemically denatured protein, or distinguish a native intrinsic disorder from chemically denatured proteins [104].

It is, however, also possible to obtain much more structural information from this class of proteins based on SAXS data. One can simplify the problem of describing such ensembles of structures as an extension of the challenge of describing complex mixtures, only in this case with an astronomically large number of individual conformations being represented in the sample. Evidently this simplification is not entirely just, but it serves to illustrate the basic idea behind ensemble modeling (Fig. 3).

In SAXS-based ensemble modeling, originally pioneered by Bernađo et al. [106] the sum of calculated scattering curves from an ensemble of conformations, picked from a very large pool of possible conformations, is fitted to the experimental curve [106]. Due to the ambiguity problem several combinations will describe the data equally well. For this reason the process must be randomized and repeated, e.g. by applying a genetic algorithm as originally advocated by Bernađo et al. in the program EOM [106], an approach also implemented in ASTEROIDS, which has the significant advantage that coupled refinement against NMR and SAXS data can be performed [107,108]. Also the size of the conformational ensemble must be refined during the procedure [109], and the statistics of the individual selections must guide the final selection. An elegant recent development in ensemble selection is the sparse ensemble selection method, which sets the upper limit of the ensemble size based on the experimental information content [110]. A representation of the overall biophysical parameters (e.g. \( R_g, D_{\text{max}} \) or inter-domain distances in multi-domain proteins) of the selected pool versus the random pool is a way of presenting the results [106,109]. Such a representation plots the distribution of conformational features that represents the experimental data, rather than just the average bulk value and hence is superior and provides more information than any average calculation. It is however not possible to determine the total number of conformations present in the sample, and the selected ensemble is representative of the overall features specifying the structure of this particular IDP under these particular experimental conditions.

Alternative approaches focus on selecting minimal ensembles, basis set ensembles or a smaller number of clusters of conformations (similar at lower resolution) that adequately fit the data [111–113]. The principal focus here is to avoid over-fitting of the data, which may happen by including too many free parameters.
when adding a larger number of conformations. The basic way of thinking is in that way orthogonal to the original idea, where it is advocated that since IDPs are weighted ensembles of a very large number of conformations, then also the advocated that since IDPs are weighted ensembles of a very large number of conformations does fit to the experimental data, this does not mean that these particular conformations are the actual and only conformations present in solution. Again, the conformations are regarded as representative for the degree of flexibility of the molecule under investigation. Not only the selection procedure, but also the generation of the initial pool varies in the different ensemble approaches available. The initial conformational pool can be random, assuring a full coverage of the 3D space [106] and may contain just the monomeric representation of the protein or also include e.g. symmetric oligomers [109], or the pool can be enriched by certain conformations, generated by the user and hence based on prior knowledge [114]. Arguably one of the most advanced method available so far for generating the initial pool (and which was originally developed for the use in NMR refinement [115]) is flexible mecano, including prior knowledge about the structural propensity for the specific bioinorganic amino acids [116]. In this latter case, the starting pool is in accordance with the genetically encoded information in the peptide chain, and hence enriched by basic knowledge, thus should be a superior starting point compared to a random or subjectively enriched pool. The recently published BE-SAXS method stands out from the rest, in that it applies a Bayesian probabilistic model for the SAXS data and generation of protein structures when fitting to the experimental scattering profiles [117]. The method has the advantage that an arbitrary number of structures can be included in the ensemble without increasing the number of parameters fitted to the data, and, particularly, the implicit advantages of probabilistic modeling of the protein structure [70,117].

No matter the implied strategy, the important point is that it is possible to use BioSAXS to select an ensemble of structures, representative for the typically occurring structural features of the sample, and hence to obtain a kind of relative structural representation. Importantly, a database of ensemble descriptions of intrinsically disordered proteins or highly flexible proteins, called pEDB, has been established [118], collecting results from the very challenging analysis of such systems. All structures are sensitive to the experimental conditions, but the adaptive structural nature is more pronounced for intrinsically disordered proteins, hence it is an extra strength that the SAXS data can be recorded under numerous experimental conditions, thereby to a higher extent exploring the conformational space of a given intrinsically disordered protein.

9. SAXS is central in hybrid methods

SAXS is, as evidenced, a highly valid stand-alone method for the analysis of even highly complex macromolecular samples. Yet BioSAXS gains even further importance when considering hybrid methods, coupling BioSAXS with orthogonal structural and bio-physical methods [119,120]. Not only is this a way of circumventing the ambiguity problem associated with SAXS-only modeling, but also SAXS offers information which is often complementary to other types of data. One of the particular SAXS features is the extensive q-range covered by the method hence SAXS contributes with information on a broad length-scale, covering from sub-nm to µm scale. A very common and useful combination of methods is MX and BioSAXS, which was already discussed above.

Evidently, numerous other methods have been coupled with SAXS analysis. One example combines atomic force microscopy and SAXS. It was shown that the protein gephyrin, part of the postsynaptic protein complex in inhibitory synapses and consisting of two domains separated by an extended unfolded domain, exists primarily as a trimer and in a mixed compact/extended state [121].

An increasingly popular combination of methods with numerous successful applications is the combination of nuclear magnetic resonance (NMR) and SAXS [122,123]. The time- and length-scales of the dynamics that are captured by the two methods beautifully complement each other, as exemplified in the study of the dynamic equilibrium between monomers and dimers of full-length capsid protein from HIV-1 [124]. This study includes a temporal description of the extensive dynamics of the N-terminal domain respective to the C-terminal oligomerization domain. NMR and SAXS data were also coupled in the de novo structure determination of the dimer of the BetV1 protein Aha1, and the interface from this solution-based structure differed from those observed in crystalllo [125]. In a study of the folding pathways of β-2-microglobulin it was demonstrated that a long-lived intermediate state has a high aggregation propensity and incorporates into both homo- and heterodimers, suggesting that excited state dimers would potentially initiate amyloid aggregation [126]. In a recent study of mutant domain N-termini of human cardiac myosin binding protein is characterized [127]. This protein includes regions exhibiting intrinsic disorder, and indeed the study of IDP is particularly suitable by the powerful combination of NMR and SAXS and several programs elegantly couple refinement from NMR and SAXS data [108,114,115,123]. A recent study of the intrinsically disordered amyloid proteins aSN and tau is a prime example of the coupling of atomic resolution information and long-range information from SAXS data [108].

10. Spatio-temporally resolved models

The development at leading synchrotron beamlines is breath-taking, and the opportunities, focused around BioSAXS, are rapidly expanding. Of particular interest are the time-resolved small and wide angle X-ray scattering (SAXS/WAXS) opportunities [133,134] emerging with the fast detectors, refined data collection protocols and advanced sample environments (e.g. pump-probe, stopped-flow, microfluidics) as already reviewed above [24–29]. The field has recently been nicely reviewed [135] and only two examples are included here: In a recent example cooperative allostery in the tetrmeric interface of wildtype and mutant haemoglobin is observed at nanosecond time resolution upon photo-induction [136]; in a different study stopped flow experiments causing a sudden pH-jump were used to resolve millisecond aggregation of mutant apomyoglobin leading via a transiently formed monomeric species to fibrils [137]. Both examples reveal the potential of the methodology, where highly complex processes including complicated mixtures can be followed at increasingly high spatial and temporal resolution.

A completely different type of resolution may be achieved if coupling the currently possible high-resolution single-particle cryo-electron microscopy (cryo-EM) analysis [138] with advanced BioSAXS analysis. The revolution in cryo-EM enables not only high-resolution de novo structure determination but also the description of mixed states, present in the sample. Here, the complementarity of solution scattering may play a future important role in aiding to validate the relevant number of structural states that should be refined from the cryo-EM data, just as the presence of the high-resolution structures, present on the cryo-EM grid, could be validated in solution. Time will show what will be possible, but it does seem fair to say that the recent and current revolutionary development in the BioSAXS field likely will continue, probably even at
an increased speed. As always, the large-scale facilities play a central role. The amazing technological developments at synchrotron beamlines pull the development in the field, enabling ever more daunting experiments in the fascinating world of structural biology.

Acknowledgements

M. H. Jensen is acknowledged for her contribution to Fig. 1. Beamline staff members at BioSAXS beamlines around the world are greatly acknowledged for their amazing contributions to the field, and the constant support and inspiration provided over the years.

References


