Development of Methods to analyze and represent Small-Angle Scattering Data from Interacting and Flexible Biological Macromolecules

Dissertation with the aim of achieving a doctoral degree

at the Faculty of Mathematics, Informatics and Natural Sciences

Department of Chemistry of Universität Hamburg

Submitted by Mikhail Kachala

Hamburg

2015

Day of oral defense: 19.06.2015

The following evaluators recommend the admission of the dissertation:

Prof. Dr. Dr. Christian Betzel, Institut für Biochemie und Molekularbiologie (Fachbereich Chemie der Universität Hamburg)

Prof. Dr. Reinhard Bredehorst, Institut für Biochemie und Molekularbiologie (Fachbereich Chemie der Universität Hamburg)

Date of the approval for publication: 23.06.2015

To my wife Ekaterina Kalininskaya

Abstract

Small Angle Scattering (SAS) is a widely applied technique in structural biology and the number of its applications is rapidly increasing due to the advances in data collection and analysis methods. Simultaneously the systems that are characterized using this technique are becoming more diverse and complex and the amount of the experimental data is growing. This leads to the necessity for further development of advanced methods for data analysis and representation. In this thesis different aspects of SAS data processing and analysis as well as applications to various biological problems are covered. The first project is the formulation of the extension of a standard SAS archiving file definition in order to accommodate various type of data, required during SAS data analysis from a scattering curve to the final models. Besides the extension itself the project includes development of the tools for its processing and applications as well as integration with small-angle scattering databases. The second project is focused on the complex interacting systems, which are difficult for analysis, yet are important for a number of applications. In order to overcome the associated issues a Monte-Carlo based method for deconvolution of form and structure factor was developed. The third project reports on the enhancements and studies the capabilities of Ensemble Optimization Method, which is widely used in analysis of structural properties of intrinsically disordered proteins (IDPs).

Increasing number of SAS users and experiments caused an upsurge in the amount of experimental data and models based on it and has led to an introduction of the SAS databases. Currently there is no possibility to exchange information between the databases resulting in duplication and incompatibility of entries, limiting opportunities for the data driven research and creating others obstacles for the SAS data users. In this work, a solution based on the use of a widely adopted Crystallographic Information Format (CIF), is developed to resolve these issues and provide the universal exchange format for the community. An extension of a tailored sasCIF format was designed, which comprehensively describes the necessary experimental information including relevant metadata for the SAS data analysis process and for the deposition into a database. Processing tools for these files were developed and are available as standalone programs and integrated to the SASBDB database allowing export and import of the data entries as sasCIF files. The update of sasCIF and development of tools to process file of this format is an important step to standardize the way SAS data is presented and exchanged. Together with

the introduction of SAS databases, it makes the method more accessible for users and promotes its application in the structural biology community.

Interparticle interactions are not rare in solution scattering and their presence makes conventional approaches of SAS data analysis not applicable. Scattering contributions arising from the interactions between particles can affect scattering curves even at relatively low protein concentrations making the determination of the distance distribution function, essential for SAS data analysis, complicated. To separate the scattering component caused by interparticle scattering (structure factor) from the scattering data containing information about the shape of the particles (form factor) a Monte-Carlo based approach was developed. The underlying idea is a simultaneous reconstruction of the structure factor and the distance distribution function by a global procedure involving random generation of sets of parameters defining these functions. The optimization of the parameters is driven by the fit to the experimental data and boundary conditions. The approach was tested on both synthetically generated and experimental SAS data and the obtained results show that it can quantify and reconstruct structure factor contributions and provide distance distribution functions in both cases.

Intrinsically disordered proteins and proteins with intrinsically disordered regions are of great interest in structural biology today and SAS has become widely used for study such molecules because of the technique ability to characterize unfolded structures in solution. The aim of the third project was to analyze the capabilities of Ensemble Optimization Method (EOM) – one of the most widely used methods for analysis of SAS data for disordered proteins. Although the current version of EOM 2.0 has been released, there were several aspects of its application to be investigated. Conducted tests have shown that EOM 2.0 is able to correctly represent properties of the unfolded proteins, resolve distinct conformations as well subpopulations of flexible structures and robust to the noise in scattering curves.

Besides data analysis methods development, several applications of SAS to biological problems (for both folded and disordered proteins) are presented in this thesis. In the course of these projects, the entire palette of data analysis methods from basic data reduction to advanced techniques such as rigid body or multiphase *ab initio* modelling was applied. The results of collaborative projects with EMBL the beamline users become a part of the studies revealing structure and properties of the various proteins.

Zusammenfassung

Kleinwinkelstreuung (auf Englisch; small angle scattering, SAS) ist eine häufig angewandte Technik in der Strukturbiologie. Aufgrund des Fortschrittes in der Datenerfassung sowie in den Analysemethoden steigt die Anzahl der möglichen Anwendungen rapide an.

Gleichzeitig sind die Systeme, die mit dieser Technik charakterisiert werden können, immer vielfältiger und komplexer, so dass die Menge an experimentellen Daten zunimmt. Dies hat zur Folge, dass die weitere Entwicklung von fortschrittlichen Methoden der Datenanalyse und -darstellung notwendig wird. In dieser Arbeit werden verschiedene Aspekte der SAS Datenverarbeitung und -analyse besprochen, sowie einige möglichen Anwendungen, die der Beantwortung vieler biologischen Fragestellungen dienen können.

Das erste Projekt in dieser Arbeit ist die Formulierung und die Erweiterung einer Definition für ein Standard SAS Archivierungsdateiformat. Hiermit werden die verschiedenen Arten von Daten erfasst, die notwendig sind, um ausgehend von der gemessenen Streukurve ein Modell zu berechnen. Des Weiteren wurden in diesem Projekt Werkzeuge entwickelt, die für die Bearbeitung dieses Formates für anschließende Anwendungen notwendig sind. Zudem wurde dieses Dateiformat für die Integration in SAS-Datenbanken erweitert. Das zweite Projekt konzentriert sich auf die Bearbeitung von biologischen Systemen, die durch komplexe Wechselwirkungen gekennzeichnet und daher nur schwer zu analysieren sind. Diese Systeme sind jedoch für eine Reihe von Anwendungen wichtig. Hier wird eine Monte Carlo Methode beschrieben, die entwickelt wurde, um die damit verbundenen Probleme zu überwinden und dennoch aus den SAS Daten den Form- sowie den Strukturfaktor zu ermitteln. Das dritte Projekt befasst sich mit neu implementierten Verbesserungen an der Ensemble Optimization Method, die zur Strukturanalyse intrinsisch ungeordneter Proteinen sehr häufig verwendet wird. Des Weiteren wurden die möglichen Anwendungsgebiete dieser Methode genauer studiert.

Die steigende Zahl von SAS Benutzern und Experimenten führte zu einer Zunahme von experimentellen Daten und Modelle, die auf SAS basieren. Dies machte die Einführung von SAS-Datenbanken erforderlich. Derzeit gibt es jedoch keine Möglichkeit, Informationen zwischen den einzelnen Datenbanken auszutauschen, was zu Mehrarbeit und Inkompatibilität der Einträge führt sowie zur Einschränkung der Möglichkeiten für datenorientierte Forschung als auch die Entstehung von weiteren Hindernissen für die SAS-Datennutzer. In dieser Arbeit

wird eine Lösung beschrieben, die auf die Verwendung des weit verbreiteten Crystallographic Information Format (CIF) beruht, um diese Probleme zu lösen und um ein universelles Austauschformat für SAS Nutzer zu schaffen. Eine Erweiterung eines maßgeschneiderten sasCIF Formates wurde für die umfassende Beschreibung der erforderlichen experimentellen Informationen optimiert, einschließlich der relevanten Metadaten für die SAS-Datenanalyseprozesse sowie die Eintragung in einer Datenbank. Bearbeitungswerkzeuge wurden für diese Dateien entwickelt und als Standalone-Programme zur Verfügung getellt. Diese können in die SASBDB Datenbank integriert werden, was den Export und Import der Dateneinträge als sasCIF Dateien ermöglicht. Die Aktualisierung des sasCIF Formates und die Entwicklung von Werkzeugen, um dieses Dateiformat zu verarbeiten, sind wichtige Schritte auf den Weg zur Präsentation und dem Austausch von SAS Daten. Zusammen mit der Einführung von SAS Datenbanken wird der Umgang mit dieser Methode für die Nutzer erleichtert und deren Anwendung in der Gemeinschaft der Strukturbiologen gefördert.

Interaktionen zwischen mehreren Partikeln in Beugungsversuch mit biologischer Probenlösung häufig beobachtet und erschwert die SAS Datenanalyse, da viele der herkömmlichen Ansätzen in solchen Fällen nicht anwendbar sind. Streusignale, die von diesen Wechselwirkungen zwischen den Teilchen stammen, können die Streukurven schon bei relativ niedrigen Proteinkonzentrationen beeinflussen, was die Bestimmung der Verteilungsfunktion der Abstände erschwert, die jedoch für die SAS-Datenanalyse entscheidend sind. Um die Streukomponente, die von solchen inter-partikulären Streuung (Strukturfaktor) stammt von den Streudaten mit Informationen über die Form der Partikel (Formfaktor) zu unterscheiden, wurde ein Monte-Carlo-basierter Ansatz entwickelt. Die zugrunde liegende Idee, ist eine gleichzeitige Rekonstruktion des Strukturfaktors und der Verteilungsfunktion der Abstände durch ein globales Verfahren, welches bestimmte Sätze von Parametern für diese Funktionen zufällig erzeugt. Die Optimierung der Parameter wird durch die Anpassung an die experimentellen Daten und Randbedingungen angetrieben. Der Ansatz wurde an künstlich generieten sowie experimentellen SAS Daten getestet. Die Ergebnisse zeigen, dass eine Quantifizierung möglich ist und Beiträge für die Strukturfaktoren rekonstruiert werden können. Zudem konnten in beiden Fällen Verteilungsfunktionen der Abstände ermittelt werden.

Intrinsisch ungeordnete Proteinen, sowie Proteine mit ungeordneten Domänen, stellen heutzutage in der Strukturbiologie eine besondere Herausforderung dar. Dabei wird SAS für die

Studie solcher Moleküle sehr oft eingesetzt, da diese Technik die Charakterisierung von entfalteten Strukturen in Lösung erlaubt. Das Ziel des dritten Projekts war es, die Möglichkeiten der „Ensemble Optimierung Methode“ (EOM) zu analysieren, welches das am weitesten verbreitete Verfahren zur Analyse von SAS Daten von ungeordneten Proteinen ist. Obwohl die aktuelle Version des EOM 2.0 freigegeben wurde, benötigten mehrere Aspekte dieser Anwendung detaillierte Untersuchungen. Durchgeführte Tests haben gezeigt, dass EOM 2.0 in der Lage ist, richtige Eigenschaften der ungefalteten Proteinen darzustellen, Konformationen sowie Subpopulationen von flexiblen Strukturen zu lösen und verlässlich in Bezug auf das Rauschen der Streukurven ist.

Neben diesen Methoden der Datenanalyseentwicklung sind mehrere Anwendungen von SAS bezüglich biologischen Fragestellungen (für gefaltete sowie ungeordnete Proteine) in dieser Arbeit behandelt. Im Rahmen dieser Projekte wurde die gesamte Palette der Methoden der Datenanalyse von Grunddatenreduktion zu fortgeschrittenen Techniken wie starre Körper oder mehrphasige *ab initio* Modellierung angewendet. Die Ergebnisse der Kooperationsprojekte mit EMBL-Beamline-Benutzern sind bezüglich der Analyse der drei-dimensionalen Struktur und den Eigenschaften der verschiedenen Proteine dargestellt.

Contents

List of figures

List of tables

List of abbreviations

1D – one-dimensional

2D – two-dimensional

3D – three-dimensional

AUC – analytical ultracentrifugation

BSA – bovine serum albumin

CIF – crystallographic information framework

DDL – dictionary definition language

DESY – Deutsches Elektronen-Synchrotron

DLS – dynamic light scattering

D_{max} – maximum dimension of the particle

FRET – Förster resonance energy transfer

FT – Fourier transform

HPV – human papillomavirus

IDP – intrinsically disordered protein

IDR – intrinsically disordered region

IFT – indirect Fourier transform

MM – molecular mass

mmCIF – macromolecular Crystallographic Information framework

MSF – modified structure factor

MX – macromolecular X-ray crystallography

NMR – nuclear magnetic resonance

PDB – Protein Data Bank

R_g – radius of gyration

SANS – small angle neutron scattering

SAS – small angle scattering

SASBDB – Small Angle Scattering Biological Data Bank

sasCIF – small angle scattering crystallographic information framework

SAXS – small angle X-ray scattering

STAR – self-defining text archival and retrieval

Acknowledgements

This thesis and the projects presented in it would not be possible without the support and contributions from many people. First of all, I would like to thank the entire BioSAXS group for fruitful discussion, useful feedbacks and great atmosphere in this three and a half years, Al Kikhney for his valuable advices in both small angle scattering and programming, my special gratitudes to Cy Jeffries, Haydyn Mertens and Melissa Gräwert for proofreading of my thesis.

Many thanks to the entire EMBL administration team for creating the perfect environment for the scientific research and helping me with every organizational issue.

I acknowledge the European Comission (the 7th Framework Programme) Marie Curie grant IDPbyNMR (contract No 264257) and the Bundesministerium für Bildung und Forschung project BIOSCAT (Grant 05K20912) for providing the fellowship for my PhD.

I would like to thank collaboration partners for working together on interesting projects: David Albesa-Jové (section 5.2 I27-PimA fusion protein); Jana Skerlova (section 5.3 CD44 MEM-85 antigen-antibody complex); Isabella Felli, Roberta Pierattelli and Eduardo Calçada (section 5.4 E7 HPV Disordered protein) and Ladislav Bumba (section 5.5 RTX domain of CyaA protein).

I thank my parents Nadezhda and Vadim Kachala for the initial impulse and constant encouragement, and my wife Ekaterina Kalininskaya for the unimaginable support.

I would like to gratefully acknowledge my university supervisor Christian Betzel and my Thesis Advisory Committee members Matthias Wilmanns and Edward Lemke for useful suggestions regarding my projects.

Finally, I owe my deepest gratitude to my supervisor Dmitri Svergun for coming up with the challenging projects, constant guidance and extremely helpful feedback. His supervision not only made this thesis possible, but also helped me to acquire knowledge and skills that are crucial for my future career.

Introduction

Investigation of structural properties of the biological macromolecules is an important task in modern molecular biology, because it is crucial for understanding of molecular mechanisms that underlie biological functions and possible associated diseases. Small-angle scattering (SAS) is a powerful technique that is used for analysis of the structure, structural changes and interactions of proteins, nucleic acids and their complexes in solution [1]. Since the first SAS experiments in 1930s [2] the method was applied for characterization of dispersed particles and later was extended to biological macromolecules [3]. Initially SAS was used to determine only basic parameters of the molecules, such as radius of gyration R_g, but development of data analysis methods has made possible extraction of information about particle shape from the scattering data [4]. An important step in SAS field was made in the recent decades with the introduction of third generation high brilliance synchrotron radiation sources, which decreased data collection time down to a few seconds, and also new neutron radiation facilities enabling meaningful biological experiments. These developments caused significant increase in number of applications of both small angle X-ray scattering (SAXS) and small angle neutron scattering (SANS) for structural characterization of biological macromolecules in solution in wide range of molecular masses from kilodaltons to gigadaltons [5]. Another reason for the increase of interest in SAS was the introduction of powerful and user-friendly data analysis and modelling methods implemented for example in ATSAS package [6-8]. The programs from this package allow rapid generation of low-resolution (1-2 nm) three-dimensional (3D) models of the particle, and the reconstruction can be performed either with no prior information (*ab initio* methods) or using high-resolution structures obtained with other techniques such as macromolecular X-ray crystallography (MX) or nuclear magnetic resonance (NMR) with a rigid body modelling approach. Quantitative characterization of flexible structures and mixtures is possible with the new methods of data analysis as well.

The advances in both instrumentation and data analysis methods increased availability of SAS for structural biologists leading to a tremendous increase in the number and diversity of applications and therefore in the amount of experimental data and complexity of models. To accommodate and disseminate the wealth of SAS data and related models SAS databases were launched [9, 10], but currently there are no opportunities to exchange data between them, and

that causes obstacles for both database maintainers and their users. To overcome this problem and to improve the access to the SAS data, the world-wide Protein Data Bank (wwPDB) small-angle scattering task force recommended to develop a standard file format for SAS data exchange that includes all the relevant information stored in the databases [11]. As part of this PhD work an extension of the existing SAS Crystallographic Information Framework (sasCIF) format [12] was designed to solve that task. To make the files of this format usable by the structural biology community appropriate processing tool (sasCIFtools) were developed and integrated into the SASBDB database [9]. The results of this project are presented in Chapter 2.

Common SAS data analysis methods are applied for the diluted solution of biological macromolecules [13], when the interparticle interactions can be neglected. However with the growing number of SAS experiments the number of cases is increasing where interparticle interactions play a significant role. In such cases the usual data analysis approaches are not applicable and interaction between particles must be taken into account. To separate the scattering contribution caused by these interactions (structure factor) from the information about the shape of the particles (form factor) a Monte-Carlo based method was developed and presented in Chapter 3.

One the most actively investigated topic in structural biology today is characterization of intrinsically disordered proteins (IDP) and proteins with intrinsically disordered regions (IDR). SAS has become the important experimental technique used for this challenging task [14], largely thanks to the development of new analysis methods able to quantitatively assess the flexibility. A major milestone of this development was the Ensemble Optimization Method (EOM) [15], and I was involved in the work on the new release EOM 2.0 [16]. Several important aspects of the method were investigated and Chapter 4 presents the analysis of the capabilities and limitations of the program. The extensive tests checked how well can EOM 2.0 represent properties of the unfolded proteins, verified its ability to resolve subpopulations in mixtures and its robustness to the noise in the scattering data.

Applications of SAXS to concrete biological problems in the frame of collaborative projects with EMBL P12 beamline users (structural characterization of folded, disorder proteins and protein complexes) are presented in Chapter 5. The data analysis approaches used in these projects include basic data reduction methods but also advanced techniques such as rigid body refinement and multiphase *ab initio* modelling. The structural parameters and models obtained

based on SAS data were integrated with the results of other methods for a comprehensive characterization of the investigated proteins.

Chapter 1. Basics of SAS

1.1. Introduction

Small angle scattering (SAS) is a powerful technique for investigation of structural properties of macromolecules and nanoparticles in solutions e.g. at native conditions. Among advantages of the technique are absence of limitations on the size of the molecule and no need in crystallization, which makes SAS measurements possible in cases when the molecule is too large for NMR or crystals are not available for MX. The resolution of the technique is about 1 nm, which allows determination of size, shape, conformation, oligomeric and folding states of molecules. Since the measurements are performed in solution the structural modifications of the studied macromolecules in response to the changes in the environment, for example temperature or pH, can be investigated as well.

The first SAXS experiments discovering the potential of the method were performed in 1930's by the French physicist Andre Guinier [2] and later the method was applied for analysis of the structural parameters of biological macromolecules in solutions [17] The recent advances in high-flux synchrotron radiation sources, neutron radiation facilities and in-house X-ray instruments made solution scattering more accessible and very popular among structural biologists. New methods of data analysis and modelling allow one to build elaborate models either *ab initio* or by incorporating high-resolution structures obtained with other methods such as MX or NMR.

In this chapter an overview of basic theoretical concepts of solution scattering, SAS experiment, data processing and modelling of biological macromolecules are presented. This work mainly relates to SAXS, and neutron scattering (SANS) is specifically discussed only when the difference between the techniques is substantial.

1.2. Solution scattering theory

SAS theory is based on the elastic scattering of X-ray photons by electrons (SAXS) or neutrons by nuclei (SANS). Elasticity means that only the photons/neutrons that do not change their energy are registered and the wavelengths of the incident and scattered radiation are equal. When, in the case of SAXS, the sample consisting of macromolecules in solution is irradiated by a monochromatic X-ray beam with wavelength λ of approximately 0.1-0.15 nm all electrons

within the macromolecule are becoming sources of secondary spherical waves [18]. As we consider only elastic scattering modulus of the initial wavevector $\boldsymbol{k}_0$ and the secondary wavevector $\boldsymbol{k}_1$ is the same ($|\boldsymbol{k}_0| = |\boldsymbol{k}_1| = 2\pi/\lambda$), but the direction is different. The difference between the two wavevectors $\boldsymbol{s} = \boldsymbol{k}_1 - \boldsymbol{k}_0$ is called either scattering vector or momentum transfer with the magnitude equal to

$$s = \frac{4\pi sin\theta}{\lambda} \tag{1}$$

where 2Θ is the scattering angle (Fig. 1).

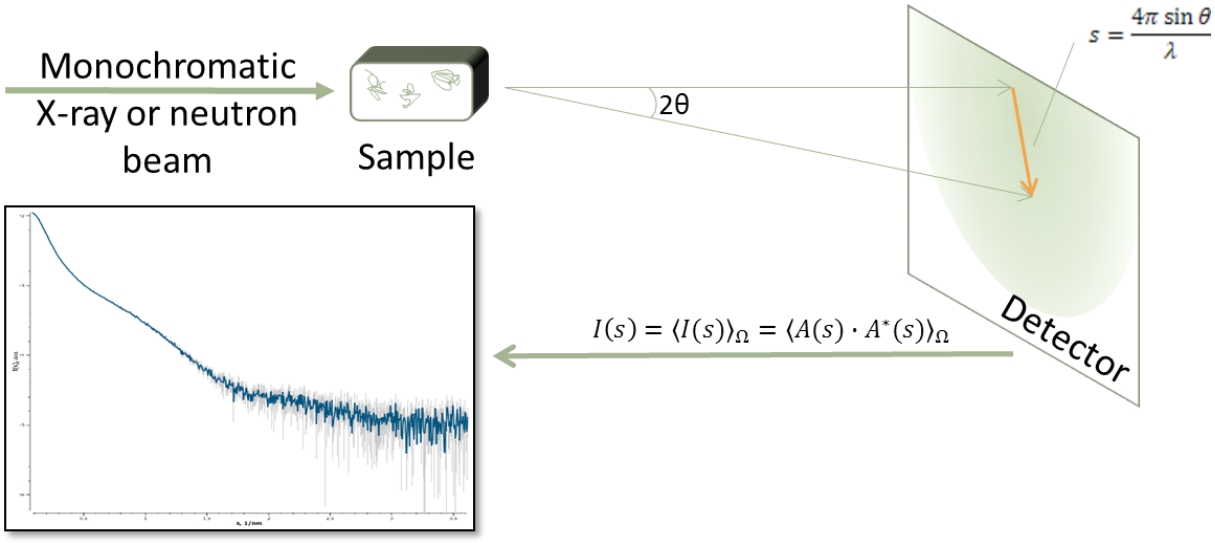

Fig. 1. Scheme of a typical SAS experiment.

To describe the scattering from the sample the scattering length density distribution $\rho(\boldsymbol{r})$ is introduced, where $\boldsymbol{r}$ is a coordinate vector (for SAXS, this is electron density proportional to the number of electrons per unit volume). In solution scattering we are interested in the scattering from the macromolecules not the solvent so the excess scattering length density is considered to be $\Delta\rho(\boldsymbol{r}) = \rho(\boldsymbol{r}) - \rho_s$, where ρ_s is the density of the solvent. The amplitude of scattering can be expressed as a Fourier transform of the excess density:

$$A(\boldsymbol{s}) = \Im[\Delta\rho(\boldsymbol{r})] = \int_V \Delta p(\boldsymbol{r}) \exp(is\boldsymbol{r}) d\boldsymbol{r} \tag{2}$$

where the integration is performed over the volume of the particle. The detectors can register only intensity but not the amplitudes, and all (randomly oriented) particles in the illuminated

volume scatter X-rays so the intensity is a product of amplitude and its complex conjugate averaged over all possible orientations:

$$I(\boldsymbol{s}) = \langle A(\boldsymbol{s})A^*(\boldsymbol{s})\rangle_\Omega \quad (3)$$

The SAXS or SANS intensity function I as a function of momentum transfer s is called scattering curve and this one-dimensional curve is used in further analysis.

1.3. SAS experiment

In a SAXS or SANS experiment besides sample itself the buffer (pure solvent) measurement must be performed. The buffer scattering is then subtracted from the sample scattering in order to get scattering only from the dissolved particles. Three main sample requirements for SAS experiments are

i. purity of the sample (95% monodisperse or better);
ii. absence of unspecific aggregates;
iii. measurements should be done at different solute concentrations.

The concentration series is used to extrapolate the sample signal to infinite dilution (zero concentration). The useful signal coming from the solute depends on the number of macromolecules in the sample, i.e. the high concentration samples have a better signal to noise ratio. However, higher concentrations can lead to more pronounced interparticle interactions (see Chapter 3) that alter the scattering at lower s values, thus hindering data analysis. Therefore, the extrapolation to infinite dilution is a crucial step in the scattering experiments and subsequent data analysis.

The unspecific aggregates, formed due to strong attractive interparticle interactions, must be avoided in the sample because they significantly change the scattering patterns and the data from such samples is not usually suitable for further analysis. Hence, prior to the SAS measurements a check for sample purity is necessary using other techniques for example dynamic light scattering (DLS), gel filtration chromatography or analytical ultracentrifugation (AUC).

The major advantage of SANS is the possibility of deuteration of the sample or solvent, which is a very powerful approach for characterizing multicomponent macromolecular complexes, however the sample preparation in this case is long, expensive and difficult [19]. SANS experiments typically require larger sample quantities (about 300 µl) and there are fewer neutron radiation sources available than X-ray ones. Nonetheless, SANS experiments are important for

structural characterization of macromolecular complexes and their parameters should therefore be included in SAS data exchange standards (see Chapter 2).

1.4. Basic data processing

1.4.1. Pair-distance distribution function

A very important information about macromolecules that can be obtained directly from the scattering curve is the pair-distance distribution function $p(r)$. The function represents a histogram of distances between volume elements within the particle weighed by their excess scattering density [18]. The distance distribution function can be calculated using the inverse Fourier transformation of the scattering curve (Fig. 2):

$$p(r) = \Im[I(s)]^{-1} = \frac{r^2}{2\pi^2}\int_0^\infty s^2 I(s)\frac{\sin sr}{sr}\,dr \qquad (4)$$

The $p(r)$ function is best obtained via indirect Fourier transform approach [20] implemented, for example in programs ITP [20] and GNOM [13, 21]. The value of r beyond which $p(r)$ is equal to zero is a maximum particle dimension (D_{max}) being one of the most important particle parameters directly determined in a SAS experiment.

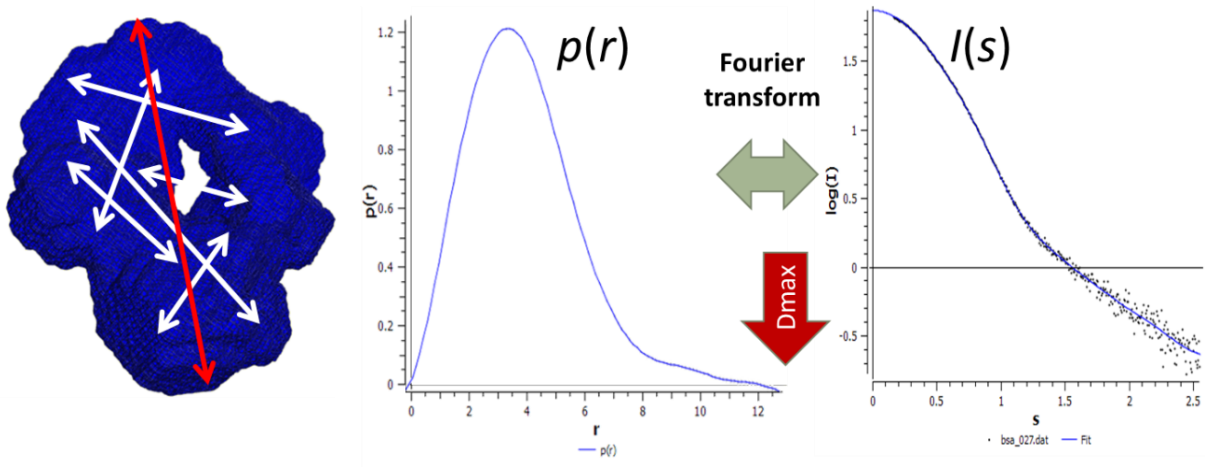

Fig. 2. Determination of distance distribution function

1.4.2. Radius of gyration R_g and Guinier region

Another important parameter that can be determined directly from the scattering curve is the particle radius of gyration R_g, a square root of the averaged squared distance from the

center of mass weighted by the scattering length density. The method of R_g determination developed on the dawn of solution scattering era is called Guinier approximation and is described at small angles ($sR_g < 1.3$) by the equation (5):

$$I(s) = I(0)\exp\left(-\frac{s^2R_g^2}{3}\right) \quad (5)$$

Alongside with R_g the forward scattering $I(0)$ can be also derived using the equation (5). The approach to determination of these two parameters lies in the use of a linear region of the so called Guinier plot ($\ln(I(s))\ vs\ s^2$), slope of which is corresponding to R_g and intersect with the y-axis yields $I(0)$.

1.4.3. Molecular mass determination

The value of forward scattering $I(0)$ is proportional to concentration and to the molecular mass (MM) of the particles in the sample. Therefore the value of $I(0)$ normalized against concentration can be used for the estimation of the molecular mass. In a typical SAS experiment, a standard protein (usually Bovine Serum Albumin (BSA) with the MM of 66 kDa) is measured separately and the molecular mass of the solute is calculated according to equation (6):

$$MM_{sample} = \frac{I(0)_{sample}}{I(0)_{standard}} MM_{standard} \quad (6)$$

where MM_{sample} and $MM_{standard}$ are molecular masses and $I(0)_{sample}$ and $I(0)_{standard}$ are forward scattering of the sample and standard protein respectively. This method is straightforward and widely applied for the molecular mass estimation, but it is sensitive to the incorrectly determined concentration.

An alternative approach to molecular mass determination is based on the hydrated particle volume (Porod volume), which is estimated from the scattering curve using the Porod's equation [22]:

$$V_{porod} = \frac{2\pi^2 I(0)}{\int_{s_{min}}^{s_{max}} s^2[I(s) - K]ds} \quad (7)$$

where K is a constant subtracted to ensure the asymptotical intensity decay to s^{-4} at high s values, integral in denominator is called Porod invariant Q and s_{max} is empirically estimated as $8/R_g$ [23]. To calculate the molecular mass in kDa the Porod volume in nm^3 should be divided by about 1.6 (with accuracy of approximately 20%) [8].

1.5. Modelling based on SAXS data

1.5.1. Ab initio modelling

The first approach for reconstruction of the 3D shape of the particles using the solution scattering data was proposed by H.Stuhrmann in the 1960s [24]. The method is based on utilization of spherical harmonics expansion and allows a rapid analytical calculation of the scattering curve from a known 3D shape represented as an angular envelope function describing the particle border [25]. With the use of trial-and-error approach and/or optimization methods the inverse problem could be solved and the envelope of the particle could be determined.

Another approach to shape analysis using reverse Monte-Carlo based bead modelling was suggested by Chacon [26]. In a program DAMMIN [27] the advantages of bead modelling were coupled with the speed of the spherical harmonics. The algorithm represents the shape in a search volume (usually a sphere with diameter equal to D_{max}) filled with densely packed beads. Each bead can be assigned either to a particle or a solvent. The procedure starts with a random configuration, which is optimized with a simulated annealing algorithm to find the structure with the scattering pattern, which has the minimal discrepancy with experimental. The looseness or lack of interconnectivity of the structure is penalized during the optimization allowing one to exclude unphysical solutions. The scattering curves from the generated models are rapidly computed using spherical harmonics, and yet faster version of the algorithm – DAMMIF – was recently introduced [28], which generates an *ab initio* model within a minute on a standard PC, thus opening the way for a high-throughput analysis of SAXS data. The expansion of DAMMIN to the modelling of multiphase objects (e.g. nucleoprotein complexes) is implemented in program MONSA [27]. In this case the beads can be assigned not only to the protein or solvent, but also to each component (phase) of the particle.

1.5.2. Rigid body modelling

In some cases, high-resolution information about the macromolecule obtained with X-ray crystallography or NMR is available prior to the SAXS measurement. If the entire structure is known SAS can be used to confirm if the molecules in solution have the same shape with the programs CRYSOL (for SAXS) [29] and CRYSON (for SANS) [30]. If the structures of components (subunits of domains) are available, then their mutual location and be determined with the automated rigid body modelling implemented in the program SASREF [31]. Similarly to *ab*

initio methods, the optimization starts from a random configuration of the components, and a simulated annealing algorithm modifies the positions and orientations of the subunits to minimize the goal function. The latter contains the discrepancy term and a penalty term introduced to avoid steric clashes and to ensure interconnectivity of the constructed models. Additional restrictions based on the information from complementary methods can be imposed during rigid-body modeling, for example, contact conditions known from cross-linking experiments.

However the complete structures of the subunits are often not available, for example when linkers are missing from the high-resolution models, and in this case the rigid body approach cannot be applied directly. To overcome the limitation, methods to combine rigid body and *ab initio* modelling were implemented in BUNCH [8] and CORAL [31]. Both programs operate similarly to SASREF, the main difference between them lying in the generation of the missing portions of the structures. BUNCH represents them as chains of dummy residues, and CORAL select the linkers form a precomputed library of native-like fragments.

1.6. Conclusion

The overview of the small-angle scattering theory, SAXS experiment, basic data processing and modelling given in this chapter shows that the technique has diverse applications allowing one to comprehensively characterize biological macromolecules at low resolution. In the next chapters, further improvements in SAS data storage, exchange, processing and modelling developed in the course of this PhD project are presented and illustrated by examples of solution scattering applications to structural characterization of proteins and protein complexes.

Chapter 2. Extension of sasCIF file format and development of sasCIFtools

2.1. Introduction

The past decade has seen a significant increase in the popularity of biological SAS to investigate the shapes of macromolecules in solution [32]. Consequently, the means to store data and models and to make these results accessible to the structural biology community has become a priority, both in terms of transparency and quality assurance [11]. SAS databanks have been recently introduced [11] to help address this issue, for example BIOISIS (www.bioisis.net at the Lawrence Berkeley National Laboratory [33]) and Small Angle Scattering Biological Data Bank, or SASBDB (www.sasbdb.org developed at EMBL Hamburg [9]). Both databases contain SAS data and models, but currently there is no possibility to exchange information between them, and this may cause problems with respect to data management, duplication and incompatible frameworks. The lack of agreed-to standards with respect to data deposition also limits and complicates the development of data mining and analysis protocols, thus creating obstacles for future data-driven research. Finally, cross-platform exchange of experimental data, analysis protocols, general experimental information and the results obtained using various instruments and radiation sources (X-rays or neutrons) is hindered by the lack of a consistent and user-friendly file structure. Discussions on this issue within the SAS community has resulted in the recommendation by the wwPDB small-angle scattering task force to develop a standard file format for SAS data exchange that includes all relevant information stored in the databases [11]. The natural candidate to fulfill this role is the sasCIF format introduced in 2000 [12] that was initially designed as a convenient and easily convertible format to exchange one-dimensional (1D) SAS data between laboratories. However, SAS databases store other types of information in addition to the scattering data including structural parameters, data-transforms (e.g., Guinier and Kratky plots), real space pair-distance distributions, volumes as well as various types of models of the particle(s) in question and their respective fits to the data. Consequently, the current definition of sasCIF has to be extended to accommodate all types of data plus auxiliary information required for SAS projects and be compatible for database deposition. The sasCIF format also has to provide convenient tools to handle and process the data, including sasCIF converters for both current and historical projects so that formatting the data and information for database deposition is a seamless process. The aim of the project described in Chapter

2 was to design the sasCIF extension according to community needs, to develop sasCIF tools as part of the ATSAS package [8], to process redefined sasCIF files and to make available sasCIF import and export options for database entries.

2.2. Basics of CIF organization

The sasCIF format is a part of Crystallographic Information Framework (CIF) [34] family of file formats, which are becoming standard across the structural biology community [35]. The CIF format was initially introduced as a general purpose data exchange format for small molecule X-ray diffraction experiments [34]. The initial CIF format, as well as its subsequent derivatives, consists of three key elements: Self-defining Text Archival and Retrieval (STAR), Dictionary Definition Language (DDL) [36] and a data dictionary for a specific domain (e.g. SAS or NMR) based on DDL (Fig. 3). Later the extension of CIF files for the description of macromolecules including the atomic coordinates of high-resolution models, mmCIF (macromolecular CIF), was developed and has become a standard format to represent crystallographic data [37]. Other CIF formats are built on similar principles and include dictionaries to describe electron microscopy structures (3DEM), NMR data (NMRSTAR), two-dimensional (2D) detector data (IMGCIF), etc.[1] For small-angle scattering, the sasCIF format [12] also uses this highly adaptable framework. The hierarchical and key elements of CIF are shown in Fig. 3.

2.2.1. STAR

The Self-defining Text Archival and Retrieval (STAR) format was developed as a machine-independent universal archive format [38, 39]. The main feature of STAR is its ability to store any type of numerical or textual data. Another important property of STAR is the self-defined structure of items, i.e. one is able to interpret file content correctly with no prior knowledge of file structure (contrary to .pdb format). A STAR file is defined as an ASCII file, which contains data organized according to STAR syntax. The syntax defines following elements:

i. A *data item* is the data value.
ii. A *data name* is the name of data item.
iii. A *loop* is a list of repeated data items used for tabular data.

[1] The full list of the CIF dictionaries can be found here: http://mmcif.wwpdb.org/dictionaries/downloads.html

iv. A *saveframe* is a collection of data items, names and loops and;

v. A *data block* is the collection of i-iv, above.

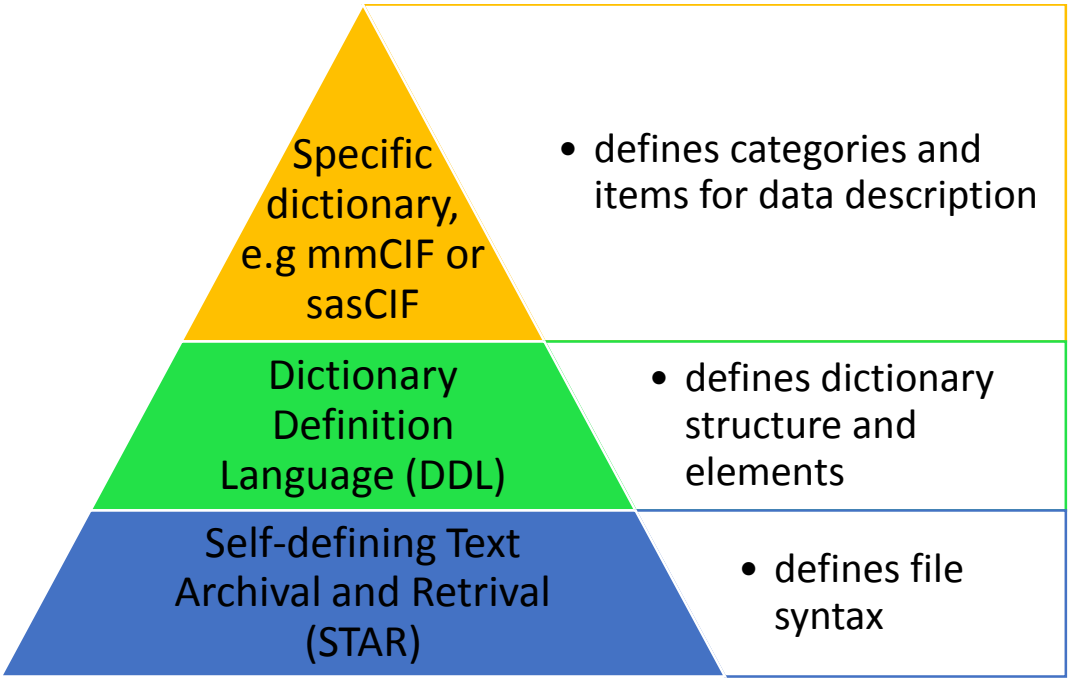

Fig. 3. Structure of CIF formats dictionary definition

The *data block* is identified with a string "data_*blockcode*", where *blockcode* is a unique code within a STAR file. A *saveframe* is identified in the similar way: "save_*framecode*" at the beginning of the save frame and by the string "save_" or next data block in the end. *Data name* is a string starting with an underscore symbol ("_") and *data item* is a string, which does not start with an underscore and preceded by the data name. The unknown or missing data values are denoted as a period "." or as a question mark "?", if they are not relevant. A *data name* must be unique within a *saveframe* or in cases where *saveframe* is not defined in a data block. A *loop* structure is preceded with the "loop_" string and consists of several lines of *data name*s followed by the *data items*. The values cannot be missing from the loop and their quantity must be the exact multiple of the number of *data item*s. The text value is determined as either a sequence of nonblank symbols, or a sequence of characters surrounded by single or double quotes, or in case of multiple lines, by semicolons. The comment lines start with sharp sign "#". The definition of the STAR format was updated in 2012 and the main improvements are the full UNICODE character set support and new of data structure containers [40], however in sasCIF only basic data structures (*data block*, *saveframe* and *data name/item*) are employed.

2.2.2. DDL

The dictionary definition language (DDL) defines the structure of the dictionaries for a specific structure domain and the basic elements of such dictionaries [35, 41]. The simplified structure of a generic CIF dictionary is shown in Fig. 4. The current version of DLL is 2.1.15 and is available here: http://mmcif.wwpdb.org/dictionaries/mmcif_ddl.dic/Index/. According to DLL each dictionary has a *name*, *version history* and *methods* applied to *data blocks*, however currently methods are not used in CIF dictionaries. As DLL is based on STAR it uses the same syntax and structures. A dictionary itself is contained within the *data block* and each definition is stored in a *saveframe* inside this *data block*. The definitions are name-value pairs, the name is the component to be defined and the value is the definition.

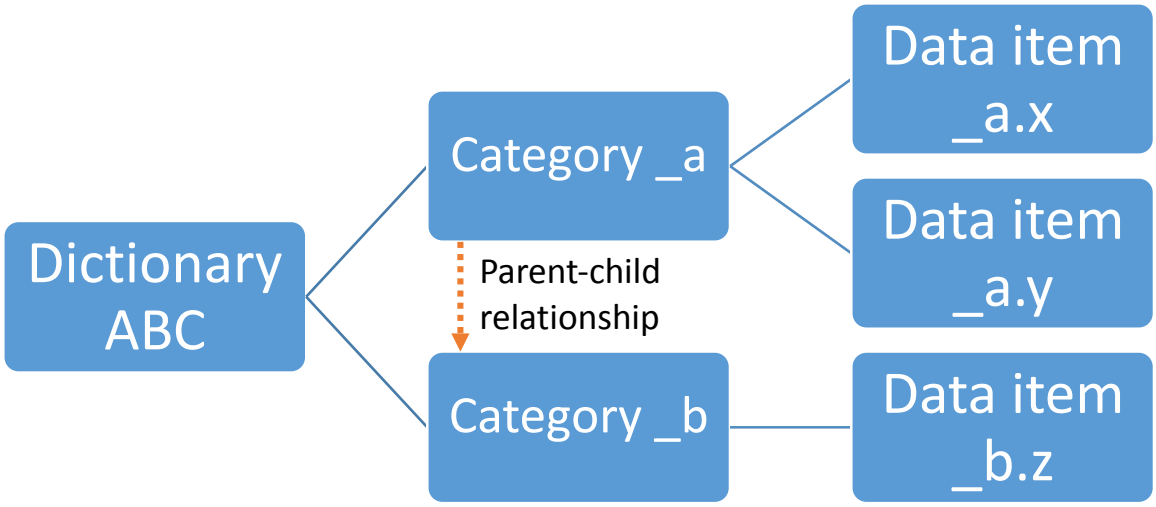

Fig. 4. Overall structure of a generic CIF dictionary

The upper organizational levels of a dictionary are category groups and categories. An example of a `sas_beam` category description is shown below. The category has a name (`sas_beam`), a description a primary key used for unambiguous identification (`'_sas_beam.id'`) and an example of the category usage. A category group consists of one or more related categories and each category can belong to several groups.

```
save_sas_beam
    _category.description
;              Items in this category give information about the beam.
;
    _category.id                    sas_beam
    _category.mandatory_code        no
    _category_key.name             '_sas_beam.id'
    loop_
    _category_group.id              'sas_group'    'beam_group'
    loop_
```

```
    _category_examples.detail
    _category_examples.case
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
-
;
    Example   - Hypothetical example to illustrate the description of a
                beam geometry.
;
;
    _sas_beam.id                          X_ray
    _sas_beam.axis_id                     1
    _sas_beam.shape                       Rectangular
    _sas_beam.width_ax                     0.2
    _sas_beam.width_eq                     0.5
    _sas_beam.dist_spec_to_detc            1.5
    _sas_beam.radiation_wavelength_id      3
;
    save_
```

The most basic element of a dictionary is a *data item*, which, similar to STAR, has a name and a value. Each data name begins with an underscore character, and the name of category and item is separated by dot. The *data item*s that belong to sasCIF extension have prefix `_sas`. A sasCIF *data item* is defined in the *saveframe* that contains its name, description and properties, such as data type or units of the value. In the example below the definition of the sasCIF extension *data item* `_sas_beam.radiation_wavelength` is given. This item is not mandatory, so the value of `_item.mandatory_code` is `no`, the type of data is `float` and the wavelength is measured in `angstroms`. The DDL provides other possibilities to define item properties, for instance give example of values or related items, but they are not widely used in sasCIF.

```
save__sas_beam.radiation_wavelength
   _item_description.description
;      The wavelength of the incident beam in Angstroms.
;
    _item.name                    '_sas_beam.radiation_wavelength'
    _item.category_id              sas_beam
    _item.mandatory_code           no
    _item_type.code                float
    _item_units.code               'angstroms'
    save_
```

Another important feature specified in the DDL is the parent-child relationships, which allows for complex data structures in the files. The relation should be specified within the parent category as well as separate *saveframe*. In the following example the parent category is `_sas_detc` and the child – `_sas_scan`, and `_sas_scan.detc_id` is a pointer to the parent. *Data items* `_pdbx_item_linked_group.link_group_id` and

`_pdbx_item_linked_group_list.link_group_id` are the index of this parent-child relation for child category.

```
#Parent-child relationship within the category
    _item_linked.child_name  '_sas_scan.detc_id'
    _item_linked.parent_name '_sas_detc.id'
#Parent-child relationship within the dedicated save frame
    loop_
    _pdbx_item_linked_group.category_id
    _pdbx_item_linked_group.link_group_id
    _pdbx_item_linked_group.label
#...
    sas_scan   3 sas_scan:sas_detc:3
#...
    loop_
    _pdbx_item_linked_group_list.child_category_id
    _pdbx_item_linked_group_list.link_group_id
    _pdbx_item_linked_group_list.child_name
    _pdbx_item_linked_group_list.parent_name
    _pdbx_item_linked_group_list.parent_category_id
#...
    sas_scan       3 '_sas_scan.detc_id'        '_sas_detc.id'    sas_detc
#...
```

The last element of CIF framework is the specific domain dictionaries that, like sasCIF and mmCIF, define the categories, *data items* and relationships among them. Each specific dictionary is defined by the means of DLL and STAR and is therefore self-describing.

2.3. Structure and content of sasCIF

Initially, sasCIF was designed to exchange 1D scattering data [12] and, similar to other CIF file formats, the sasCIF dictionary is based on STAR and DDL described above. Every sasCIF file must follow the definitions of categories, *data items* and their relations provided in the dictionary. The sasCIF dictionary has following structure:

1. Dictionary description. Name, general information and version of the dictionary.
2. Dictionary history. The history of dictionary updates including versions and dates (changelog).
3. Subcategories definitions. Currently only one subcategory “vector” is defined within the sasCIF dictionary and used for description of axis parameters in `sas_axis` category imported from imgCIF.
4. Definitions of data types, e.g. “code”, “text”, “float”, etc.
5. List of units used for numeric values.

6. Definitions of category groups. Name, description and categories, which belong to the group.
7. Definitions of categories and data items (Main part) with all the categories and data items definitions.
8. Parent-child relations descriptions.

Key elements components of this structure, in particular the definitions of data types and category groups are described below.

2.3.1. Data types and categories

Definition of data types. The sasCIF files have the capacity to store diverse information and data types. In the dedication section of the dictionary, for each data type, its code, primitive code (number or character), allowed characters (as regular expressions) and description are given. Table 1 outlines the types of data used in sasCIF and examples of their use.

Table 1. Data types used in sasCIF.

Code (name)	Primitive code	Description	Example
code	character	code item types/single words	`1`
ucode	character	code item types/single words (case insensitive)	`24`
line	character	char item types/multi-word items	`"Bos taurus"`
text	character	text item types/multi-line text	`;MAHTVAGES GSAHLKDPD ;`
int	number	the subset of numbers that are the negative or positive integers	`8`
float	number	the subset of numbers that are the floating numbers	`10.950`
yyyy-mm-dd	character	Standard format for CIF dates	`2012-09-17`

Categories. The original sasCIF format included 1D scattering data as well as information about the beam, detector, and sample. In 2007 the standard was extended to encompass categories describing experimental parameters (`sas_scan`) as well as the coordinate system used in the experiment (`sas_axis`), for example the type of scaling used for the momentum transfer,

s. In Table 2, an overview of the categories in the 2007 version of the sasCIF dictionary (v 0.4) is presented.

Table 2. sasCIF 0.4 categoreis

Category	Description
`sas_scan`	Parameters of the experiment
`sas_scan_intensity`	Scattering intensity
`sas_detc`	Information about detector
`sas_sample`	Sample information
`sas_beam`	Properties of the beam
`sas_axis`	Coordinate system used in the experiment

Since its original introduction [12] sasCIF has been integrated into a number of applications, including those at the DUBBLE CRG beam line of ESRF [42] and in the ATSAS software package [6]. However, with the growing quantity of SAS data and the introduction of SAS databases and publication standards [43] the need for sasCIF dictionary updates is pressing, and this has been indicated by the wwPDB Small-Angle Scattering Task Force [11]. In the following sections updated extensions to sasCIF and the tools necessary for sasCIF processing are presented.

2.4. Updates to the sasCIF dictionary

As shown in Chapter 1 various types of data are used during the course of SAS data analysis. The 2D scattering patterns, registered by the detector, are reduced to 1D scattering profiles, from which the real space pair-distance distribution functions, *ab initio* and hybrid models, model fits against the data and general metadata about the sample, experimental parameters, etc., are generated. As the radial averaging procedure and therefore translation of 2D scattering pattern into 1D curve is nowadays done automatically at most beam lines and all further manipulations are performed with 1D curve it makes sense to store the 2D image only in the dedicated imgCIF format. Every other piece of data is present in the SAS databases, but not all of them are available in the CIF file definition, so the previous version of sasCIF dictionary (0.4) had to be extended to accommodate all relevant data.

To determine which kinds of data should be added to sasCIF, the contents of the SASBDB [9] and BIOISIS [10] databases were examined. In general, it was found that key information types could be included into the sasCIF 0.5 dictionary to address the needs of the present and future SAS user communities:

- Results of SAS measurements, including concentration or contrast variation series information.
- Results obtained for the standardization of scattering intensities, including absolute scaling or scaling relative to secondary standards.
- Guinier analysis and probable real space distance distribution functions.
- Experimental structural parameters derived from the data; R_g, $I(0)$, D_{max} and, importantly, MM information extracted from various approaches (from $I(0)$ and concentration, Porod volume, *ab initio* volumes, etc).
- Description of the macromolecular samples, including sequence information.
- Information about the sample environment, including supporting solvent composition, temperature, pH, contrast, etc.
- Spatial (3D) models and calculated model scattering profile fits to the data, with statistical reporting of data-model discrepancies.
- Author names, affiliations and publication information.
- Cross-database links and information, including Uniprot (for proteins), PubMed and the protein databank (PDB).

A relational diagram of the updated version of sasCIF is presented in Fig. 5. In order to accommodate all required information, new categories and *data items* have been added to the sasCIF dictionary, including categories adopted from the mmCIF dictionary, while some existing sasCIF categories have been supplemented with new *data items*, and others completely redesigned as new categories. In parallel with these changes, updates of parent-child relationships between the categories have been implemented.

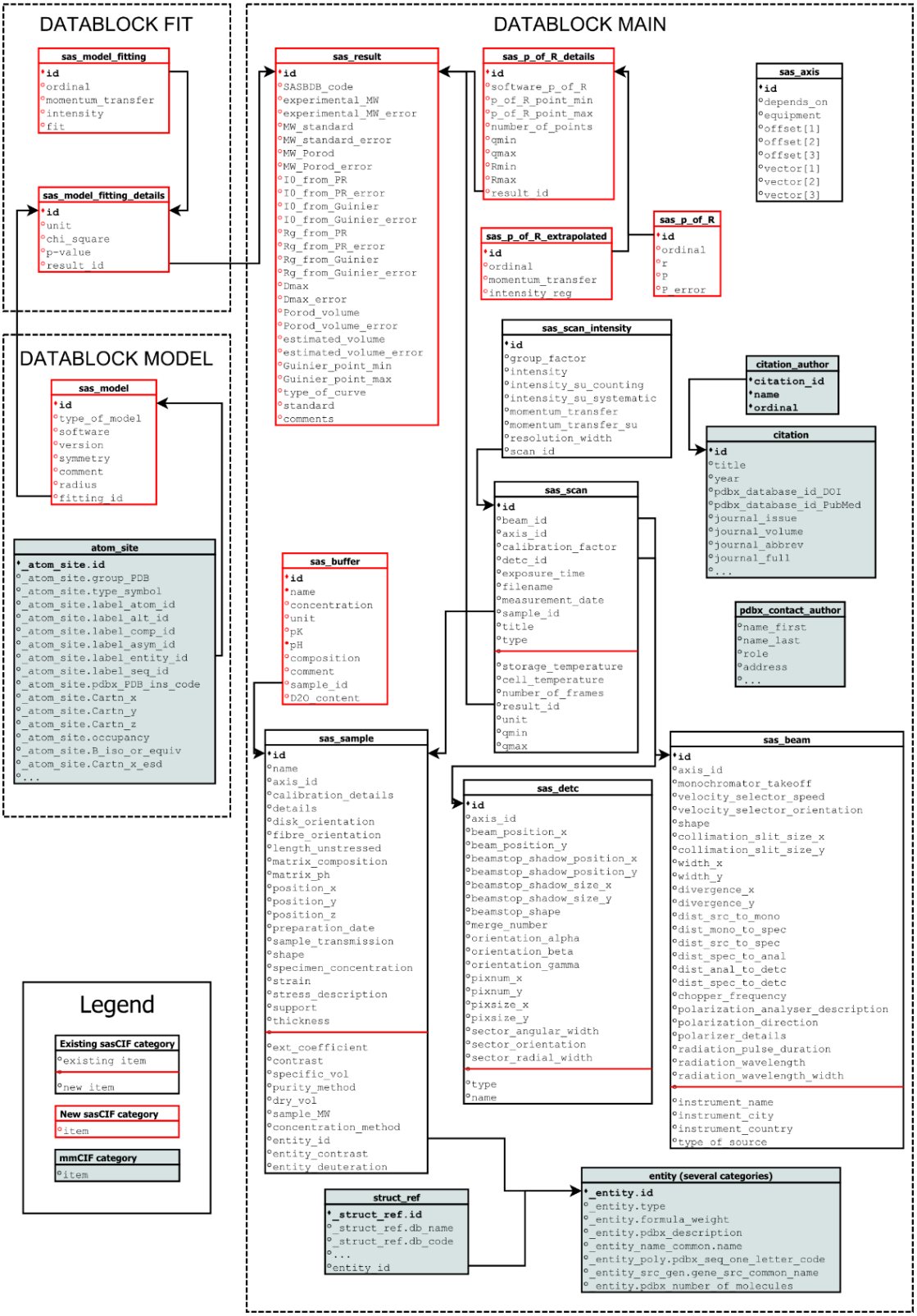

Fig. 5. A relational diagram of the updated sasCIF dictionary as developed in this chapter. The data items existing in the previous version(s) are shown in black boxes, while new categories are shown in in with their associated items. Items from mmCIF dictionary are in boxes with grey background.

2.4.1. New, expanded and additional categories in sasCIF

New categories that have been incorporated and written into sasCIF.

`sas_result`: this category contains information about the results of the measurements and the experimental errors for each parameter, e.g. radius of gyration, molecular mass, maximal dimension, etc.

`sas_p_of_R_details, sas_p_of_R, sas_p_of_R_extrapolated`: a set of categories that describe the pair-distance distribution ($p(r)$ vs. r.) The three-element structure of the category is based on the structure of the output file provided by GNOM [13], i.e., input intensities, reciprocal-space fit to the intensities and the extrapolated reciprocal space fit to zero angle that generates the distance distribution via the indirect Fourier transformation. Distance distribution information is divided into three categories, because according to CIF syntax the loop_ structures such as $p(r)$ and extrapolated intensities must be stored in separated categories.

`sas_model_fitting_details, sas_model_fitting`: The fitting details of a calculated model against the data and the statistical reporting of the fit, such as a χ^2 value, is stored apart from the fitted calculated model curve.

`sas_model`: The category describing the properties of the refined model used to interpret the SAS data. The category includes type (*ab initio* or rigid body model), the software used to build/refine the model, model symmetry (P1, P2, etc.) and, in case of *ab initio* models, the radius of individual dummy atoms used to represent the shape of the particle.

`sas_buffer`: The buffer/solvent description, including small-molecule components, concentration, pH, etc.

Existing sasCIF categories that have been expanded:

`sas_sample`: New *data items* have been added to describe sample component macromolecule UV-Vis absorption extinction coefficients, X-ray contrasts, partial specific volumes, dry volumes, molecular mass, additional methods used to assess data quality (e.g., size exclusion traces, gel electrophoresis images) the sample name and sample concentration. As sasCIF is a standard format for SANS as well as SAXS data, the neutron contrast and level at which a macromolecule is isotopically labelled with non-exchangeable deuterium are also included. Finally, the `sas_sample` category contains a pointer to `_entity` categories describing the properties of the molecules that constitutes the sample.

sas_detc: *Data items* for the name and type of detector were added, for example photon-counting detector or CCD detector.

sas_scan: The experimental parameters were extended to include temperature data (sample storage and during data collection), the number of data frames taken to compile a final dataset, the units of momentum transfer (*s* or *q* in inverse angstroms or inverse nanometers) and the experimental *s* (or *q*) range.

sas_beam: The new *data items* contain information about the beam line, its name, and geographical location, type of source and wavelength.

Categories from mmCIF used in sasCIF files:

atom_site: Atomic coordinates of spatial models.

enitity, _entity_name_common, _entity_poly, _entity_src_gen: Categories used to describe molecules in the sample.

strucutre_ref: References to external databases, in this case UniProt (for proteins).

citation,_citation_author: Publication information, including cross-links to PubMed.

2.4.2. New parent-child relationships in sasCIF.

The introduction of new categories into the new sasCIF file system requires updates of the parent-child relationships that are presented in Table 3:

Table 3. Parent-child relations between for introduced categories

Parent	Child	Type of relationship
sas_sample	sas_buffer	One-to-one
sas_sample	entity	One-to-many
entity	struct_ref	One-to-one
sas_result	sas_scan	One-to-one
sas_result	sas_p_of_R_details	One-to-one
sas_p_of_R_details	sas_p_of_R	One-to-one
sas_p_of_R_details	sas_p_of_R_extrapolated	One-to-one
sas_result	sas_model_fitting_details	One-to-many

Parent	Child	Type of relationship
sas_model_fitting_details	sas_model_fitting	One-to-one
sas_model_fitting_details	sas_model	One-to-many
sas_model	atom_site	One-to-one
citation	citation_author	One-to-many

2.4.3. Category groups in the sasCIF dictionary and data block structure

In order to make user navigation in the CIF structure easier, categories in the dictionary are deliberately grouped. Each category in the sasCIF dictionary is a member of sas_group as they all relate to SAS data. Other groups are shown in the Table 4.

Table 4. Description of sasCIF category groups.

Group	Description	Members
beam_group	Categories that describe the properties of the beam.	sas_beam
detector_group	Categories that describe the properties of the detector.	sas_detc
fitting_group	Categories that describe the fitting of theoretical models' scattering to the experimental data.	sas_model_fitting sas_model_fitting_details
intensity_group	Categories that describe the intensities	sas_axis sas_scan sas_scan_intensity
model_group	Categories that describe the models	sas_model
result_group	Categories that describe the results of the measurement	sas_result sas_p_of_R sas_p_of_R_details sas_p_of_R_extrapolated
sample_group	Categories that describe the properties of the sample.	sas_sample sas_buffer

The data blocks in a sasCIF file are not described in the dictionary, but they are defined by the dictionary structure. A category can be present in a data block only once, so in cases when several models are produced that fit against one SAS profile each model must be stored

in separated data blocks. As seen from Table 3 there are four categories with one-to-many relationships in the dictionary. Two of them, `entity` and `citation_author`, can be described with the *loop* structure, while the other two (`sas_model_fitting_details` and `sas_model`) demand dedicated data blocks (`FIT` and `MODEL` correspondingly) in the file for each model or fit. Every other category is unique within the frame of a file and is stored in the `MAIN` data block. The final data block structure of the new sasCIF file format is shown in Fig. 6.

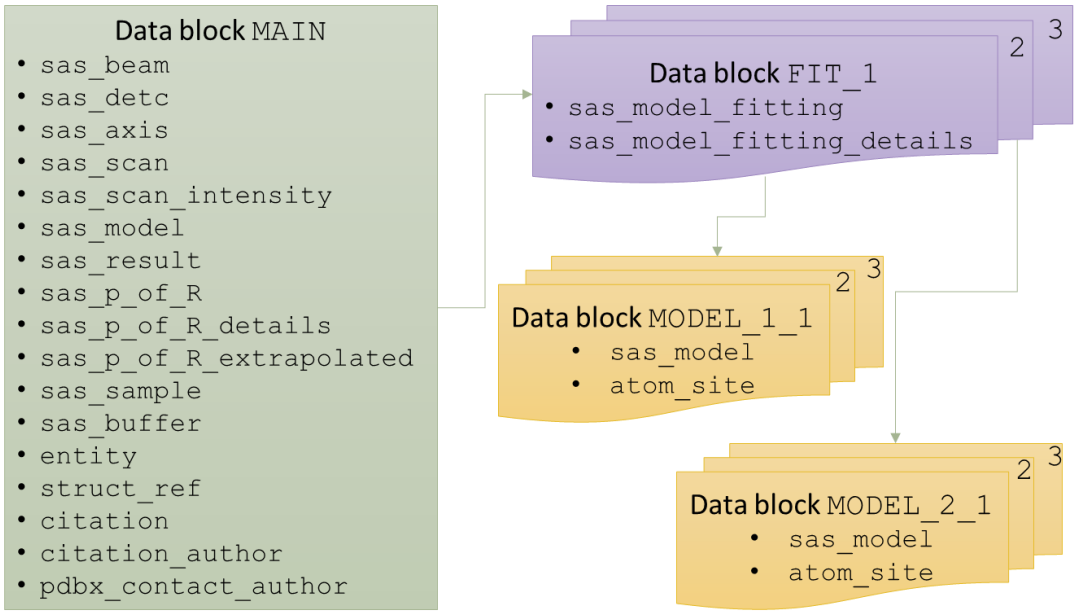

Fig. 6. Data block structure of sasCIF files.

2.5. sasCIFtools: a set of programs for processing sasCIF files

The update of sasCIF dictionary is an important step towards efficient organization and accessibility of SAS data and associated experimental information encompassing the aspects of data analysis and sample quality assurance. To make the use of the sasCIF files user-friendly, tools for sasCIF processing as well as integration into existing software and services is crucial. In the following sections, sasCIFtools for data processing and format conversion as well as the implementation of sasCIF support for the SASBDB database are described.

2.5.1. A description of data types for sasCIFtools and software implementation

The purpose of sasCIFtools is to provide the means for data conversion into sasCIF using existing ATSAS software and to enable addition and extraction of SAS information from commonly used and historical file formats. As discussed above, sasCIF files include several types of data:

- Metadata about the sample, experimental conditions, results of experiment, etc.
- Scattering data.
- Probable real space pair-distance distribution(s).
- Spatial models.
- The fit of calculated scattering intensities from spatial models against the experimental data.

The software modules of the ATSAS suite have different input requirements and output file structures, depending on the operations required for the particular analysis (e.g., GNOM produces probable pair-distance distributions; DAMMIN, dummy-atom bead models and fits; CRYSOL, high-resolution spatial model fit assessments, etc). Except for metadata, the types of data recorded and the output file structure is dictated by each individual software module. The aim of sasCIFtools is to read in these software-dependent outputs and collate them into an interpretable CIF format, containing all pertinent information for an entire experiment and data analysis process. Here, we consider main ATSAS output file types and their relations with the sasCIF data items.

The scattering data

The scattering data that ATSAS software typically deals with are primarily.dat files. The .dat files contain three columns consisting of the scattering vector s, scattering intensities $I(s)$ and experimental errors $\sigma(I(s))$. In some cases, the files store parameters of the experiment (e.g. temperature, number of frames, sample concentration) in the form of name-value pairs. Thus, when the scattering data is to be transferred to the sasCIF file it should be stored in the `sas_scan_intensity` category and properties in `sas_scan`, `sas_sample` and others. The detailed relationships between the .dat file content and sasCIF categories are shown below in Table 5.

Table 5. Correspondence between .dat file parameters and sasCIF data items

.dat file parameter	sasCIF data item
Scattering vector	`_sas_scan_intensity.momentum_transfer`
Scattering intensity	`_sas_scan_intensity.intensity`
Experimental error	`_sas_scan_intensity.intensity_su_counting`

.dat file parameter	sasCIF data item
Sample code	`_sas_sample.name`
Sample description	`_sas_sample.details`
Concentration	`_sas_sample.specimen_concentration`
Cell temperature	`_sas_scan.cell_temperature`
Storage temperature	`_sas_scan.storage_temperature`
Exposure time	`_sas_scan.exposure_time`
Timestamp	`_sas_scan.measurement_date`
Beam center position (x and y)	`_sas_detc.beam-position-x,.beam-position-y`
Wavelength	`_sas_beam.radiation_wavelength`

Real space pair-distance distributions (p(r) vs. r)

One of the primary steps in SAS data analysis is the calculation of the real space distance distribution function. In the ATSAS package [8] this is achieved using GNOM, that includes several interfaces: the GNOM software itself [13], its command line version DATGNOM and via a GUI interface in PRIMUSQT [7]. The GNOM output, the .out file, contains parameters automatically refined by GNOM that optimize the indirect Fourier transformation calculation, i.e., the regularized and smoothed reciprocal space fit of $p(r)$ vs. r to the scattering data, the fit extrapolated to $s = 0$, the real space distance distribution between $0 < r < D_{max}$ and the extracted structural parameters from the distribution (R_g and $I(0)$). The distance distribution is stored as a three-column table (the distance, r, the probable value of $p(r)$, and the error in this value, $\sigma p(r)$). The associated fit of $p(r)$ in reciprocal space against the scattering data is displayed as a five-column table (s, the experimental $I(s)$, $\sigma(I(s))$, the regularized fit of the intensities against the experimental data and extrapolated regularized intensity to $s = 0$). In sasCIF, the content of the .out file is stored in three separate categories: `sas_p_of_R` is used for the distance distribution, `sas_p_of_R_extrapolated` for the extrapolated scattering curve and `sas_p_of_R_details` for other information extracted from the .out files. It should be noted that the extrapolated curve is presented only with momentum transfer values and extrapolated regularized values, as other data could be reconstructed from the scattering curve and supporting information (this reconstruction, i.e., from sasCIF to .out, is considered in the section about the cif2out tool, which creates a .out file from the sasCIF file.)

Spatial (3D) Models

Spatial models (*ab initio*, rigid body or ensembles of models) refined against, or used to describe, SAS data are stored as .pdb files. The most reasonable approach to handle 3D spatial coordinates is to employ a dedicated mmCIF format to transfer information into sasCIF. The category in mmCIF that contains atomic coordinates is `atom_site` and the correspondence between data items in this category and .pdb fields is provided by the mmCIF developers.

Spatial model fits against SAS data

The scattering calculated from three-dimensional (3D) models is routinely used during model refinement and at the completion of the refinement calculations data-model discrepancies are reported, usually as a reduced χ^2 value, or more recently using a *p*-value obtained from correlation map analysis [44]. Files containing the model fit information (.fir for *ab initio* models; .fit for high-resolution model fitting) are presented in a three-column format (momentum transfer, intensity and calculated model scattering). Unlike .out files, the .fit or .fir files do not store scaling or cropping information and thus the original experimental scattering intensities may not always be able to be reconstructed from the fit files alone. In sasCIF, the fits of spatial models against the data are presented by two categories: `sas_model_fitting` containing *s*, *I*(*s*), and the fitted model scattering, and `sas_model_fitting_details` for information about experimental units, reduced χ^2 and *p*-values. The experimental errors are stored with the original scattering data in the `sas_scan_intensity` category.

2.5.2. sasCIFtools: External and internal libraries

The sasCIFtools is a set of programs to add data from the above data types to sasCIF as well as to extract these data types and save them as separate files, i.e., to convert SAS data files and .pdb files to sasCIF and *vice versa*. To perform this task, libraries for reading and writing of the relevant files have been developed. The library used for the input and output of sasCIF files is PDBeCIF provided by EMBL-EBI [45], while for ATSAS file processing the saxsdocument library is used that was initially written as part of saxsview project at EMBL-Hamburg [46]. However, both of these libraries have been additionally modified to process sasCIF files using the updated data type requisites as described in this chapter. The tools are implemented as

Python scripts and require installed ATSAS package [8]. The individual libraries of the new sasCIF package are described below.

PDBeCIF library is designed to read and write mmCIF files. The library is pure Python, has no external dependences and is distributed under the Apache 2.0 License. Therefore, the package is easy to modify and flexible in application. Although PDBeCIF was initially developed to process mmCIF files, it does not impose restrictions on specific categories or data items. Consequently, the library can be used with any format from the CIF family including sasCIF. The CIF structures are read and stored as nested Python dictionaries, i.e., each CIF file is a dictionary of data blocks, each data block is a dictionary of categories, and each category is a dictionary of data items with item names as keys.

According to STAR, the order of structures in a CIF file, i.e. the order of categories in a data block, is irrelevant, and so is the order of the items is a generic Python dictionary. However, to make the sasCIF files more readable for users the order of elements should be predictable. To achieve this, the PDBeCIF package was modified for sasCIFtools by using an Ordered Dictionary structure instead of a generic one. This structure saves the items in the order they were assigned and therefor are determined within the individual sasCIFtools programs.

saxsdocument is a part of saxsview project [46] that provides input/output and plotting libraries for SAS data files (.dat, .out, .fir and fit). The library is available for C, Fortran and Python, but in the latter case it is implemented as read-only. To provide an option for writing SAS data files with Python, an internal library **writesaxsdoc** was developed for sasCIFtools. It allows writing properties for name-value pairs and tabular data, e.g. column-format scattering profiles.

cifutils is an additional internal sasCIFtools library that contains support functions for all sasCIFtools, for example, arguments handling.

2.5.3. Adding data to sasCIF files: Inserting tools

All sasCIF tools used to insert ATSAS output files into sasCIF have a similar interface where each has one argument – the file containing data to be added to a sasCIF file – plus input and output sasCIF file options. If no output file is specified, the data is added to the input file. If no input file is given, the output file contains only the data added. If neither of options are provided, then the output sasCIF file has the name of the data file with a .sascif extension.

The experimental scattering profile is the most basic and essential piece of information for SAS data analysis. The **dat2cif** tool has been designed to add experimental .dat data to sasCIF files. The tool reads the scattering curve and the properties from a .dat file using the saxsdocument library and either adds the data to the first data block of an already provided sasCIF file or creates a new sasCIF file. The .out file containing pair-distance distribution is read by **out2cif** that inserts the distance distribution (r, $p(r)$, $\sigma p(r)$) and the associated information (s, $I(s)$, σI(s), regularized fit and extrapolated regularized fit) as well as the structural parameters that, when combined, are required to reconstruct the distribution from the resulting sasCIF file. For inserting the atomic coordinates obtained from spatial models, e.g., from a .pdb file, the module **pdb2cif**, was developed using the `atom_site` category from mmCIF in the sasCIF dictionary. The correspondence between .pdb fields and mmCIF data items is straightforward as defined at the PDBx/mmCIF Dictionary Resources website[2]. Finally, the insert tool **fit2cif** handles the insertion of .fit or .fir files into sasCIF files. The fit2cif tool is similar to dat2cif expect for χ^2 value parsing. In some older .fit files χ, and not χ^2, is reported and therefore the value needs to be first located in the .fit or .fir file and standardized.

2.5.4. Accessing data from sasCIF files: Extracting tools cif2

All sasCIF extracting tools (except cif2all, see below) have a similar interface. The user specifies the input sasCIF file as an argument and the individual outputs of each extraction tool are written to the same folder of the sasCIF file. The **cif2dat** tool extracts s, $I(s)$, σI(s) from sasCIF along with experimental parameters (s-units) and experimental information (sample storage and data collection temperature and, if applicable, the number of data frames used for averaging). The data, parameters and information is written to a .dat file using the internal writesaxsdoc library, with the experimental information written to the file header and footer. The names for the output files are constructed according to the following template: `<sascif_filename>.dat`.

The extraction of an .out file from sasCIF to reconstruct the probable pair-distance distribution and the regularized reciprocal-space fit of $p(r)$ vs. r to the data is not trivial. In some instances information regarding how the extrapolated fit to the data was obtained to perform the indirect Fourier transformation to obtain $p(r)$ vs. r is not available. The reason for this is that

[2] Available at http://mmcif.wwpdb.org/docs/pdb_to_pdbx_correspondences.html

automated data manipulations performed by GNOM [13] may be required to optimize calculation performance, e.g., in cases when number of points in the input data is too large, GNOM will re-bin data with a larger angular step that reduces the number of data points for the calculation. As a result, the Δs increment of the experimental data and that used in the .out file may differ from each other. To resolve this issue, the intensities of the .out file must be translated back to the same Δs as the initial scattering curve. In the **cif2out** tool, Δs rescaling is performed in the following way. First, the same interval of points is selected using information about *s*-ranges stored in sasCIF data items `_sas_p_of_R_details.qmin` and `_sas_p_of_R_details.qmax`. Then the number of *s*-points is compared and if it is not the same in both pieces, a re-binning is done. This procedure is performed using DATREGRID program[3], which is a part of ATSAS package and was developed in the course of this project. In cif2out, DATREGRID is called in a template mode that is used to match the momentum transfer axis of the two scattering curves in .dat and .out. In this case, the template is the extrapolated curve from the `_sas_p_of_R_etrapolated` category and the input file is the curve from `_sas_scan_intensity` category. The output file with the correct *s*-axis intervals is made by DATREGRID and is read using saxsdocument library and added to the sasCIF-extracted .out file. The file is written with writesaxsdoc library and has name `<sascif_filename>.out`.

Spatial models consisting of atomic coordinates are extracted from the sasCIF files with the **cif2pdb** sasCIFtool, which writes data from the `_atom_site` category onto a standard .pdb file. The calculated scattering from a model and its fit against the data are extracted by **cif2fit**, which writes to the output .fit file the scattering profiles and reporting statistics (reduced χ^2 and *p*-value). As the sasCIF file may contain several models and fits, each in its own data block, the individual models and fits are saved onto separate files that have more complex filename templates reflecting the correspondence between the models extracted from sasCIF and the model fits. For fit files, the naming convention follows: `<sascif_filename>_FIT_<fit_id>.fit;` while for .pdb files: `<sascif_filename>_FIT_<fit_id>_MODEL_<model_id>.pdb`.

In addition to the information that can be stored in data files, sasCIF also contains information about the properties of the sample, solvent, parameters of the experiment, publication records and other meta-information. To extract meta-information, a **cif2sub** tool was designed

[3] Full description is available at http://www.embl-hamburg.de/biosaxs/manuals/datregrid.html

that writes all metadata onto a text file `<sascif_filename>_submission.txt` as name-value pairs. The metadata values written in this file are chosen to match those requested in the submission process to SASBDB.

Finally, the tool **cif2all** extracts all data from a sasCIF file at once. **cif2all** calls the corresponding functions from the other sasCIF extraction tools and saves the individual output files either in the current folder or in a specified directory (using the `-o` option.) For example, the following command will create output files in a folder called SASDA85:

```
python cif2all.py -o SASDA85 SASDA85.sascif
```

2.6. Integration of sasCIFtools in the SASBDB database

SASBDB [9] is at present a standalone data bank, which is designed to readily join a federated system of databases for biological SAXS and SANS data [11]. The information in SASBDB contains experimental data, structural parameters, distance distributions, modelling results and model fits plus additional information about the authors, sample environments and beam line information. SASBDB is a relational MySQL database, where each entry corresponds to an experiment (containing a single scattering curve or a collection of curves taken on a specific construct) and is identified by a seven-letters alphanumeric code starting with *SAS*.

In order to import data from the SASBDB in sasCIF format we first needed a way to extract information from the database directly. For this purpose the Python library MySQLdb [47] was used and a **sql2cif** tool was developed. The library establishes connection with the MySQL database and processes the SQL queries. MySQLdb fetches either one or all entries from the database that satisfy a query. For convenience and for further manipulations, the fetched results are stored in one nested Python dictionary that consists of table dictionaries where each element of those dictionaries is a field name–field value pair. The sql2cif tool translates the data from that dictionary to the sasCIF-like Python dictionary as specified by the PDBeCIF library. Each sasCIF data block (MAIN, FIT and MODEL) and separately `_pdbx_contact_author` and `_citation` categories are separately processed using dedicated functions in sql2cif, and this provides additional flexibility of the use with other sasCIFtools.

The correspondence between the structure and field names of SASBDB differs from the structure and data items of sasCIF. As both SASBDB and the extended sasCIF format are still

under development, it is helpful to keep this correspondence flexible, so a correspondence dictionary was developed. The dictionary is a plain text file (bdb2cif.dic) with the correspondence specified for the tables and fields of SASBDB. Each table has a dedicated block in the dictionary, which starts from the line with the keyword `TABLE` and the table name. Within the table block the correspondences between SASBDB fields and sasCIF data items are stated in the key-value format. The key (a SASBDB field) is separated from the value (a sasCIF data item) with the colon, while more-than-one values (one filed to be written in several data items) are separated with commas. A comment lines start with sharp sign ("#"). In the example below the part of the bdb2cif.dic for SASBDB table `molecule_molecule` is shown.

```
#Description of the molecules in the sample
TABLE molecule_molecule
id: _entity.id, _entity_name_com.entity_id, _entity_poly.entity_id,
_entity_src_gen.entity_id, _sas_sample.entity_id
molecular_type: _entity.type
long_name: _entity.pdbx_description
short_name: _entity_name_com.name
sequence: _entity_poly.pdbx_seq_one_letter_code
organism: _entity_src_gen.gene_src_common_name
mw: _entity.formula_weight
uniprot_code: _struct_ref.db_code
uniprot: _struct_ref.details
oligomerization: _entity.pdbx_number_of_molecules
```

The correspondence dictionary is read from bdb2cif.dic file by sql2cif and stored as a Python dictionary. In case of any changes in either SASBDB or sasCIF they can be easily included into bdb2cif.dic without modifications of the sasCIFtools code.

Although most of the fields from SASBDB can be directly translated to sasCIF data items, some fields still require pre-processing. Among such cases is a description of the oligomeric state of a macromolecule or complex, which, at present, has to be converted from word (e.g. dimer) into a number (2). Another example is the primary sequence of the macromolecule, e.g., protein sequence, presented in one-letter FASTA format in SASBDB, while sasCIF employs a plain one-letter code without comments. More complicated pre-processing is performed for the citation information, which is saved as plain text in the database and has to be parsed and stored as several data items in sasCIF (journal name, volume, issue, pages, etc.)

sql2cif as such does not have a user interface; instead, the user connection to the database is implemented in a **code2cif** tool, which manages the data export from SASBDB to a sasCIF file. Its operation schematically illustrated in Fig. 7A starts with the user providing an input

SASBDB code. Then, code2cif calls sql2cif and retrieves all the information relevant to the entry with this code. Part of the information is directly added to the sasCIF-like Python dictionary and the location of data files is communicated to other sascCIFtools called by code2cif. These tools add the scattering curve, distance distribution, models and model fits to the same dictionary. Finally, this dictionary is written onto the output sasCIF file using the modified PDBeCIF library. The name of the file consists of the SASBDB code and ".sascif" extension, e.g. `SASDAG7.sascif`. The user may choose between two modes of operation for code2cif: i) to create a sasCIF file for a specific databases entry (with –c option) or; ii) to create a sasCIF file for every entry in the database (–a option). The former option is used when code2cif is called from the web interface of SASBDB and the latter is useful for data mining as it allows access to all entries in the database in a standard format.

The inverse process of importing data to the database is shown in Fig. 7B. The import procedure is fully managed by the **cif2all** tool, which reads the input sasCIF into the Python dictionary and then calls other tools described above. Currently, the submission of the entries to the database is performed manually. Therefore, the sasCIFtools only provide the necessary files, but do not write any information directly to the database.

sasCIF is capable of accommodating all required types of SAS information and in future it could be used both as an archival file format and also as for SAS project management as a file that accumulates data during course of data analysis project (essentially, a sasCIF 'on-the-fly' record of data processing and analysis). A possible way that on-the-fly sasCIF could be implemented is presented in Fig. 8. In the first step, the scattering data, relevant experimental and selected structural parameters (R_g, molecular mass, etc.) are added to a sasCIF file automatically e.g., using an analysis software pipeline like SASFLOW [48]. After that, the results obtained from p(r) vs. r, spatial modelling and other metadata (authors, beam line, etc.) are incorporated. Finally, the sasCIF file can be submitted to a SAS database to be accessible for the community.

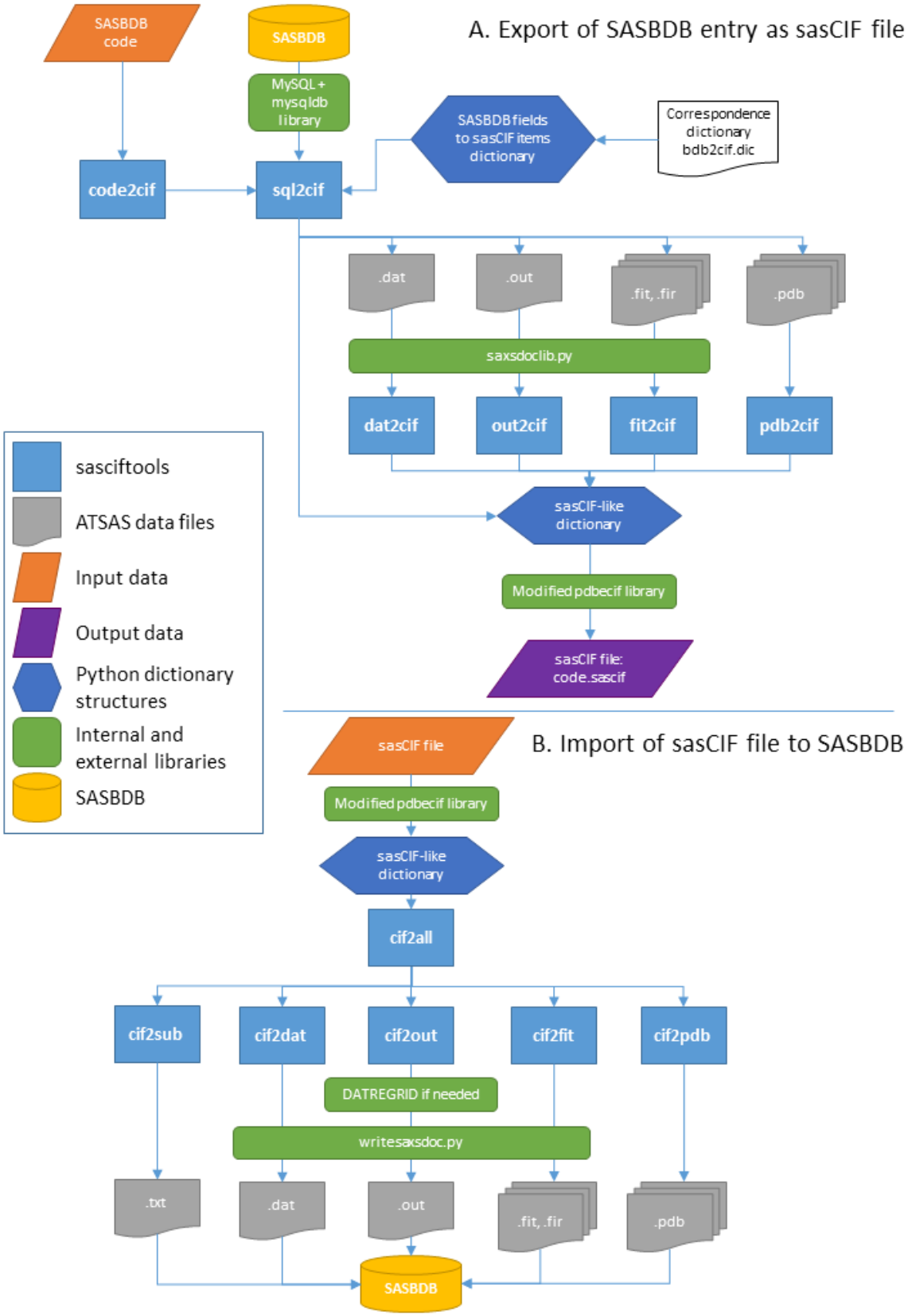

Fig. 7. Integration of sasciftools with SASBDB. A. Export from the database. B. Import to the database

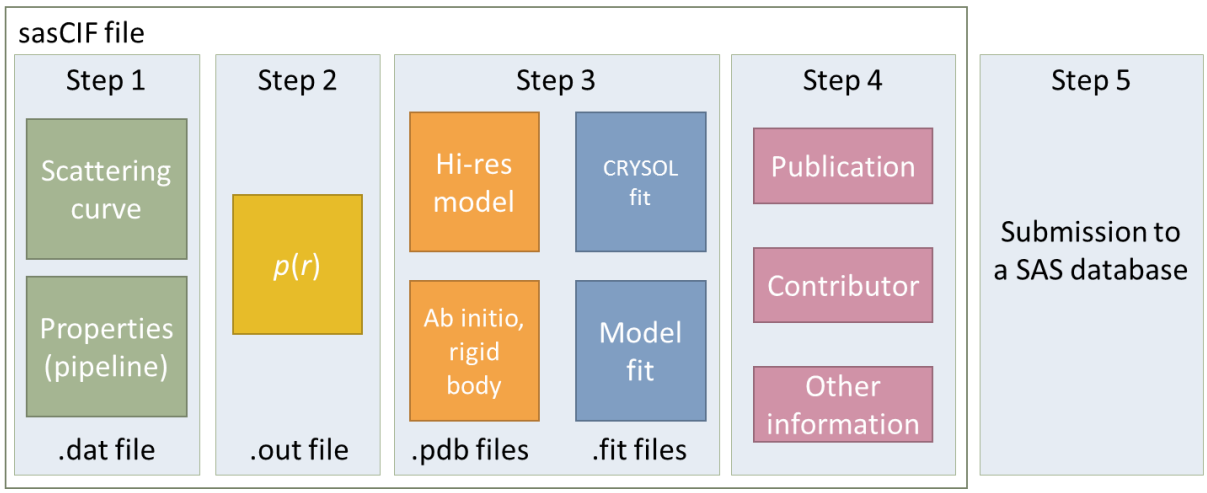

Fig. 8. Schematic use of sasCIF as a project file for SAS data analysis

2.7. Conclusion

The update of sasCIF and development of tools to process sasCIF files is an important step to standardize the way SAS data is presented and exchanged. Together with the introduction of SAS databases, sasCIF makes data organization and management more accessible for users and further promotes the application of SAS in the structural biology community improving the overall quality assurance and standards of the method.

Chapter 3. Monte-Carlo based approach for deconvolution of form and structure factors

3.1 Introduction

The determination of the pair distance distribution function, $p(r)$, is a crucial step in SAS data analysis of macromolecules in solution. The $p(r)$ provides real space model-independent information about a particle's shape [49] that can be used to reconstruct the volume and low resolution structure of molecules in three dimensions using *ab initio* methods [27]. Traditionally, the calculation of the pair distance distribution function is done via applying an indirect inverse Fourier transformation of the scattering profile as implemented in the Indirect Fourier Transformation (IFT) method [20] or GNOM program [13]. These methods work best when the SAS data are acquired from sufficiently dilute samples such that the particles do not experience interparticle interactions. If the latter are present the calculation of distance distribution function is no longer straightforward and before proceeding with analysis intra- and interparticle scattering must be deconvoluted. The intraparticle contribution to the total scattering $I(s)$ is called the form factor $P(s)$ and defined by the shape of the particles, while interparticle scattering is described by structure factor $S(s)$ and caused by their interactions [50]. Upon decoupling $P(s)$ and $S(s)$, the $p(r)$ function can be calculated from the form factor by the inverse Fourier transformation. An approach of simultaneous determination of form and structure factors is implemented in the Generalized indirect Fourier Transformation (GIFT) method [50-52], which is rather complicated and not publically accessible. The aim of the project presented in this chapter was to develop a Monte-Carlo based method for the deconvolution of the form and structure factors for a system with repulsive interparticle interactions and integrate this approach with the publically accessible GNOM software package.

3.2 Form and structure factors

3.2.1 Repulsive interparticle interactions

Scattering contributions arising from the interactions between particles may affect the scattering profiles even at relatively low sample concentrations. The impact of scattering arising from repulsive interference in case of globular particles can easily be spotted by visual analysis

of the experimental data; repulsive interactions cause a decrease in the scattering intensity at very low-*s* values followed by a distinctive interaction peak (maximum) [50] (Fig. 9A), This peak can be more or less pronounced depending on the particle shape, concentration and interaction strength. The intraparticle scattering can be described by the form factor *P*(*s*) (Fig. 9B) and interparticle – by structure factor *S*(*s*) (Fig. 9C). The measured scattering profiles are proportional to the product of these two factors so that, in general, the scattering intensity for any isotropic system of randomly tumbling particles in solution can be expressed as

$$I(s) \propto P(s)\,S(s) \tag{8}$$

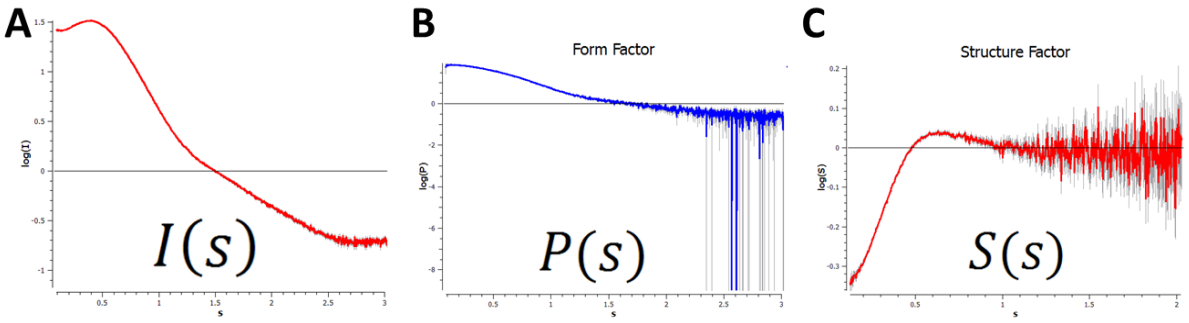

Fig. 9. Example of scattering from a sample with repulsive interparticle interactions. A. Experimental total scattering profile (I(s) vs s). B. Form factor scattering contributions, P(s). C. Structure factor scattering contributions, S(s).

If the solution is very dilute, the interactions between particles are negligible and the structure factor *S*(*s*) = 1 for all *s*. One way to experimentally eliminate structure factor influence in the data is to perform several measurements of the same sample with different solute concentrations and extrapolate the series to zero concentration. However, the structure factors are not linear and extrapolation may not provide an accurate assessment of P(s) and thus reconstruction of *p*(*r*). To obtain a *p*(*r*) function that is not influenced by repulsive interparticle interference effects the form and structure factors must be deconvoluted simultaneously.

3.2.2 Form factor and distance distribution function

The distance distribution function *p*(*r*) can be obtained by inverse Fourier transformation of the form factor *P*(*s*). However, this inverse problem is ill-posed, meaning that the solution is not unique or stable, that is why more complex methods have been developed based on reconstruction of *p*(*r*) function according to Equation (4):

$$P(s) = \int_0^{D_{\max}} K(s,r)p(r)dr \tag{9}$$

where D_{max} is the maximum dimension of the particle and the $K(s,r)$ is the kernel of the Fourier transformation:

$$K(s,r) = 4\pi \frac{\sin(sr)}{sr} \tag{10}$$

The general idea is to determine $p(r)$, which upon inverse Fourier transformation yields the experimental form factor.

The first numerical method based on this idea was Indirect Fourier Transformation (IFT) applicable for both X-ray and neutron scattering [20]. It considers the $p(r)$ function as a linear combination of the finite (20 to 30) number of *B*-spline functions. The expansion coefficients of the functions in linear combination are determined in such a way that the corresponding form factor (obtained from the Fourier transformation of $p(r)$ function) fits the experimental data. The use of the *B*-splines allows obtaining smooth solutions without additional restrictions, but to ensure that there is no oscillation in the solution a stabilization routine is required.

Another approach is employed in the GNOM software [13, 53]. The ill-posed nature of the problem is addressed by Tikhonov regularization [54] where an integral second derivate is applied as stabilizer. This way the smoothness of the solution (i.e. of the p(r) function) is also ensured. To determine the regularization parameter, i.e. the weight of the stabilizer, six perceptual parameters is used estimating the quality of the solution:

i. *oscillations*, the correct solution should be smooth and without oscillations;
ii. *systematic deviations*, the residuals in reciprocal space should have varying signs;
iii. *discrepancy* between experimental and calculated scattering curve;
iv. *stability*, how much does the solution change with the change of the regularization parameter
v. *positivity*, i.e. the share of non-negative points in $p(r)$ function;
vi. *validity in central part*, requiring the largest $p(r)$ values are the middle of $[0;D_{\max}]$ interval.

The final quality estimator is calculated as a weighted average of all criteria and used to determine the optimal regularization parameter value. GNOM is successfully applied for $p(r)$ calculations in many cases, but if one tries to use it for a scattering curve with pronounced

interparticle effect, not taking into account structure factor, the solution would be distorted. Example of such solution is shown in Fig. 10 in magenta and compared with the correct solution in red: a substantial part of $p(r)$ function is negative, which in most cases does not make physical sense for SAXS. Another problem is a substantial overestimation of the maximal dimension of the particle when determined by automated procedure implemented, for example, in DATGNOM tool [7]. This program chooses the D_{max} in a way that provides high quality of solution estimated by GNOM as well as smooth behavior of the $p(r)$ function, when it goes to zero. In some cases of interparticle interactions these conditions are met only when the D_{max} is defined as the second point with $p = 0$ ($r > 0$), which is an overestimation.

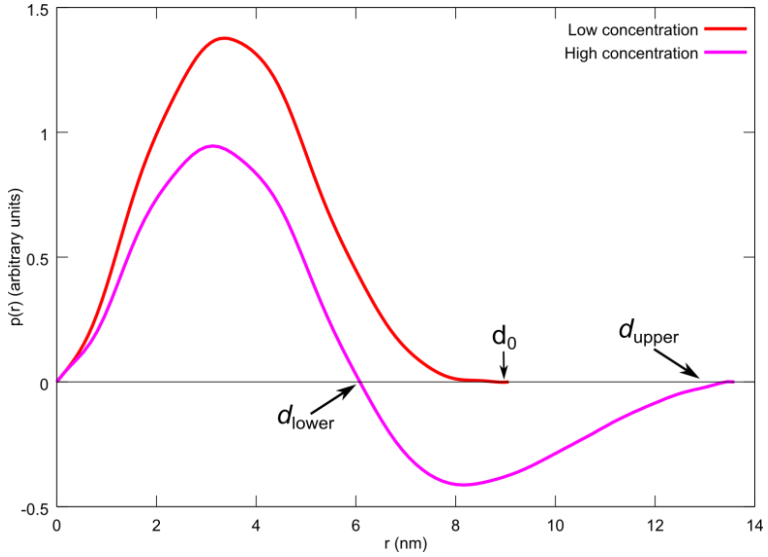

Fig. 10. Distance distribution function calculated by GNOM for low (red) and high (magenta) concentration of BSA without taking into account structure factor.

3.2.3 Structure factor and its approximation

There are several approaches to calculate structure factors both experimental, for example [55], or theoretical, for example based on solutions of Ornstein–Zernike equation [56]. Only for a handful of systems the analytical solutions for structure factors are available and a solution of hard spheres is one of them [57]. The hard sphere structure factor is determined by Perkus-Yevick approximation and is one of the most widely applied [58, 59] to model interparticle repulsive interactions. The expression for the hard sphere repulsive structure factor is calculated according to equations (11-13) as:

$$S(s, R_{hs}, v) = \frac{1}{1 + \frac{24vG(R_{hs}s)}{R_{hs}s}} \tag{11}$$

In equation (11) R_{hs} is the radius of hard spheres, v is their volume fraction in the solution and function $G(A)$ is defined as:

$$G(A) = \alpha \frac{\sin A - A\cos A}{A^2} + \beta \frac{2A\sin A + (2 - A^2)\cos A - 2}{A^3} + \gamma \frac{-A^4\cos A + 4[(3A^2 - 6)\cos A + (A^3 - 6A)\sin A + 6]}{A^5} \tag{12}$$

where

$$\alpha = \frac{(1 + 2v)^2}{(1 - v)^4}, \qquad \beta = -6v\frac{\left(1 + \frac{v}{2}\right)^2}{(1 - v)^2}, \qquad \gamma = \frac{v\alpha}{2} \tag{13}$$

Structure factors for more complex systems are also available in the literature [49]. However in this project Percus-Yevick approximation is used to describe the structure factor for simple globular macromolecules, e.g., proteins in solution, as it provides both computationally simple and reasonable starting point to model repulsive particle interactions. The Global Evaluation Technique [50] is an alternative approach of structure factor extraction also based on the Percus-Yevick approximation In this method the problem is solved by series of approximations of the $p(r)$ function and structure factor parameters. The same approach, but with a wider range of systems, including polydisperse solutions and rod-like particles was implemented in GIFT by the same authors [52], which was later modified by adding Boltzmann simplex simulated annealing optimization [51]. Another approach is implemented in IFTc method, where the structure factor is determined via real space functions [60]. However, none of these methods is accessible to the public in terms of a simple-to-use software package so the development of a new approach is desired.

3.3 Monte-Carlo based approach for deconvolution of form and structure factors

3.3.1 Problem formulation

As shown above, the pair distance distribution function $p(r)$ can be obtained from the inverse indirect Fourier transformation of $I(s)$ [13, 20]. Using equation (9) the scattering intensity (equation (8)) may be represented in the following way:

$$I(s) = NS(s)\int_0^{D_{\max}} K(s,r)p(r)dr \tag{14}$$

where N is number of particles, K is a Fourier transform (FT) kernel and $D_{\max}$ is the maximal dimension of the particle. The equation (14) is formulated for continuous functions and variables, however for computing $p(r)$ from scattering data, it is necessary that the computations take into account a predefined discretization step as experimental scattering intensities $I_{\exp}(s)$ are measured using discrete values of s over a finite s-range ($s_{\min}$-$s_{\max}$). Therefore, the integral equation 14 can be reformulated in a discrete form using:

$$\mathbf{I} = N\mathbf{S}(\mathbf{Kp}) \tag{15}$$

where **I**, **S** and **p** are vectors (lists of values) of the total intensities, the structure factor and pair distance distribution, respectively, while **K** is a Fourier transform operator matrix. Each element of the matrix is calculated by equation (10) for the specific values of s and r. If the number of s-points recorded for the intensities spanning $s_{\min}$-$s_{\max}$ is L, then the length of the vector **I** is L and so is the length of **S**. The dimension of vector **p** is defined by the number of points used for the description of p(r), and if distance distribution function is represented by M values the vector **p** consists of M values. The product of the form factor **Kp** and structure factor **S** in vector notation is an element-by-element multiplication, so the dimensions of the FT operator matrix **K** are defined by the number of points used in both the **I** and **p** i.e., L by M.

In order to calculate $p(r)$ from an experimental dataset without the influence of repulsive interparticle interference effects, i.e., to extract $S(s)$, we need to calculate the vectors **S** and **p** that satisfy equation (15) while also minimizing the discrepancy between the calculated $I(s)$ derived from **S** and **p** and the experimental scattering intensities, $I_{\exp}(s)$. The point-to-point comparison of similarity between the calculated and experimental intensities recorded at each point in s, i.e., $I_{\exp}(s_i)$ and $I(s_i)$, respectively, are assessed using the reduced χ^2 test :

$$\chi^2 = \frac{1}{L-1}\sum_{i=1}^{L}\left[\frac{\mu I(s_i) - I_{\exp}(s_i)}{\sigma(s_i)}\right]^2 \tag{16}$$

where L is the number of points in the scattering curve, μ is a scaling factor (proportional to N), and $\sigma(s_i)$ are the experimental errors. In discrete form the problem can be formulated as a minimization of the functional $T[\mathbf{S}, \mathbf{p}]$:

$$T[\mathbf{S}, \mathbf{p}] = \left\|\mathbf{I}_{\exp} - \mathbf{SKp}\right\|_{\chi^2} \tag{17}$$

However, the formulated problem remains ill-posed and in order to solve it numerically the Tikhonov regularization is applied [54, 61] where additional restrictions are imposed on the solution, for example the requirement for smoothness of the $p(r)$ function. The new target function to minimize, T_{reg}, will include a χ^2 and a stabilizer Ω (in this case discrete equivalent of integral second derivative of $p(r)$ function) with a weighting term, α:

$$T[\mathbf{S}, \mathbf{p}] = \left\| \mathbf{I}_{\mathrm{exp}} - \mathbf{SKp} \right\|_{\chi^2} + \alpha\Omega[\mathbf{p}] \tag{18}$$

To determine the **S** and **p** that minimize the functional (18) an algorithm based on simulated annealing approach was developed.

3.3.2 Description of the new algorithm

The main aim of the new algorithm described here is to simultaneously determine **S** and **p** and optimize both vectors using simulating annealing approach against the target function $T[\mathbf{S}, \mathbf{p}]$. An outline of the iterative procedure is shown in Fig. 11 and presented in more detail below. Currently the algorithm is implemented as Octave script and can be run on any operation system that has GNU Octave 3.6 [62] or higher, with package *optim* and ATSAS software suite [8] installed.

The key steps for optimizing of **p** and **S** are the initial calculation of $p(r)$ from the experimental data followed by an iterative refinement of the real-space pair distance distribution function free from the effects of repulsive interactions, i.e. to model and then remove $S(s)$ contributions from the scattering intensities. At each step of the optimization process, the $p(r)$ calculation is performed using GNOM [13]. The latter program starts with fitting the experimental data using default regularization parameters without any assumptions about the structure factor. In the case of repulsive interparticle interactions the pair distance distribution will display negative $p(r)$ values at long vector lengths bounded by two intercepts $p = 0$, d_{lower} and d_{upper}. These two values can then be used to approximate the possible range of the maximum dimensions of the particle, D_{max}, used for the second step in the iterative procedure. In this case the D_{max} is estimated to be between d_{lower} and d_{upper}, and as the simulated annealing optimization does not depend on the initial value of D_{max}, it is simply taken to be $(d_{lower} + d_{upper})/2$.

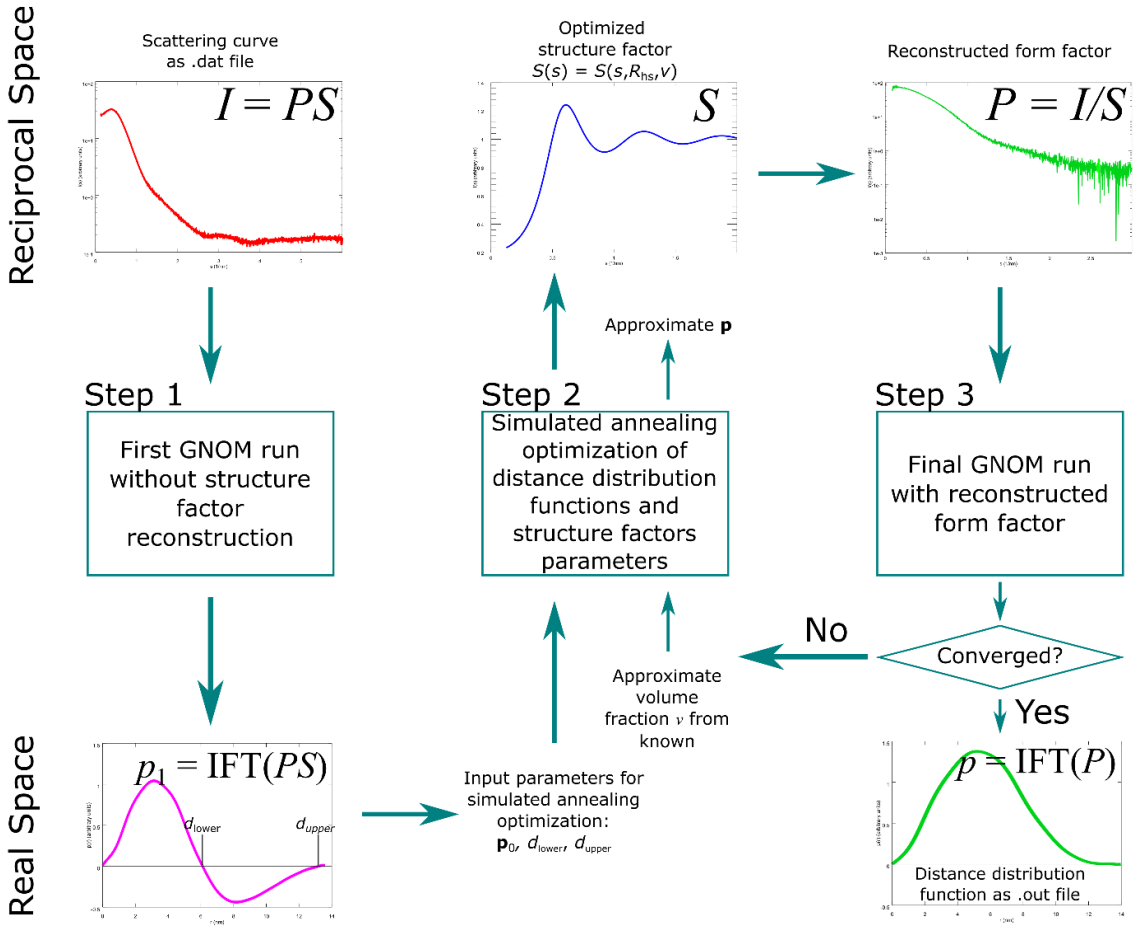

Fig. 11. Scheme of algorithm of structure factor and distance distribution function determination

Each component of the vector **p** is generated randomly as a bin in a histogram. The number of points is defined before the calculation and is chosen to provide both speed and precision. The limits in which the values of **p** can change is also determined beforehand based on the output of first GNOM run. The lower limit is set to zero, because negative values does not have physical sense for SAXS from single-component systems, and the upper limits are taken to be twice the estimated values obtained on previous iteration of algorithm.

The structure factor vector **S** is calculated according to known approximations for the repulsive structure factor $S(s)$ based on the Percus-Yevick approximation for hard spheres [57]. According to this approximation, the structure factor is defined by two parameters: i) The R_{hs} the radius of hard sphere (i.e. radius of interaction between particles) and; ii) the volume fraction v, which depends on the solute concentration:

$$S(s) = S(s, R_{hs}, v) \tag{19}$$

Thus, to determine the approximate structure factor, we use R_{hs} and v as parameters for the simulated annealing optimization. The volume fraction v can be estimated from the concentration of the macromolecules in solution by dividing concentration (in g/l) by the density of the sample (also in g/l). To determine the limits for the radius of hard spheres we used GNOM [13] for calculating the distance distribution function without taking into account structure factor. The distribution function affected by repulsive $S(s)$ contributions (Fig. 11, *magenta*) has negative values where two first points of $p = 0$ (for $r > 0$) d_{lower} and d_{upper} can be used for definition of lower and upper limits of the average distance between particles. The lower limit is set as half of the d_{lower} and the upper as $2d_{upper}$.

The simulated annealing optimization performed as the second step of the algorithm is aimed at minimizing the discrepancy between the input experimental scattering data $\mathbf{I}_{exp}$ and the intensity calculated based on the obtained distance distribution function and structure factor ($\mathbf{I} = \mathbf{SKp}$) by varying components of vector $\mathbf{p}$ and parameters R_{hs} and v defining structure factor $\mathbf{S}$. The structure factor $\mathbf{S}$ obtained as the result of the optimization is used to calculate the form factor $\mathbf{P}$ by dividing experimental intensity $\mathbf{I}_{exp}$ by $\mathbf{S}$. The obtained form factor is saved as a scattering curve and used as input for GNOM (step 3), which calculates the distance distribution function ($\mathbf{p}_{GNOM}$) – the final solution in each iteration. These three steps are repeated in the iterative manner with gradual decrease of limits of the parameters to find the converging solution.

The regularization is not performed during simulated annealing step and distribution function $\mathbf{p}$ obtained after this step may look oscillating. However the smoothness condition is imposed in the final GNOM run, so the resulting distance distribution function $\mathbf{p}_{GNOM}$ does not have oscillations.

To check the convergence of the iterative process we need to define a metric that will reflect the quality of the solution. The attributes of a poor solution are persistent occurrence of negative values in the calculated distance distribution function, an overestimated D_{max} and low quality estimation by GNOM. The first attribute is a sign of an underestimation in the structure factor and the second points to its overestimation. The quality estimate obtained from GNOM is a cumulative estimator of the solution fidelity, which takes into account the stability of the solution, oscillation of the distance distribution function, etc. [13]. To formulate the metric based on these values a series of tests with various input parameters was performed in four steps.

1. Synthetic scattering curves from solutions of interacting spherical particles with 10 nm radius and with volume fractions from zero to 0.2 with step 0.02 were generated using the Perkus-Yevick approximation for structure factor.
2. A set of structures factors was calculated with volume fractions range from 0 to 0.2 (step 0.005) and radius from 5 to 15 nm (step 0.2), in total of 2091 structure factors.
3. Each simulated curve was divided by each structure factor and the obtained "form factors" were used as input for GNOM.
4. Resulted distance distribution functions were assessed by the number of negative points, D_{max} and quality of the solution and correct solutions were compared with the wrong ones.

The results for volume fractions 0.02, 0.06, 0.10 and 0.20 are shown in Fig. 12, the correct solutions are marked by circles and based on their properties a cumulative estimator was proposed. The estimator is calculated by following rules:

i. if the share of negative points in the resulting vector **p** at a given iteration is more than 10% of all points, the estimator is equal to 0 (i.e., a bad solution).

ii. if the share of negative points is less than 10%, the estimator is calculated by equation (11), where α is a scaling parameter. The value of this parameter should be is chosen in such a way that the first summand is on the same scale as the quality estimation, for example α can be equal to initial estimation of $D_{max} = (d_{lower} + d_{upper})/2$:

$$\frac{\alpha}{D_{max}} + Quality \tag{20}$$

Finally, this estimator can be used as the convergence criterion to check if the final solution is suitable. The condition for the termination of the iterative process is reached if the change in estimator is less than 1% compared to the previous iteration. To ensure that the algorithm will not enter the infinite loop the number of iterations is limited to 100.

The basic algorithm has been tested and works well for the system of hard spheres described here. However, when applied to experimental scattering data the reconstruction of the structure factor and $p(r)$ was less accurate. To make its operation more stable and precise several improvements of the algorithm were implemented, which are considered in the next section.

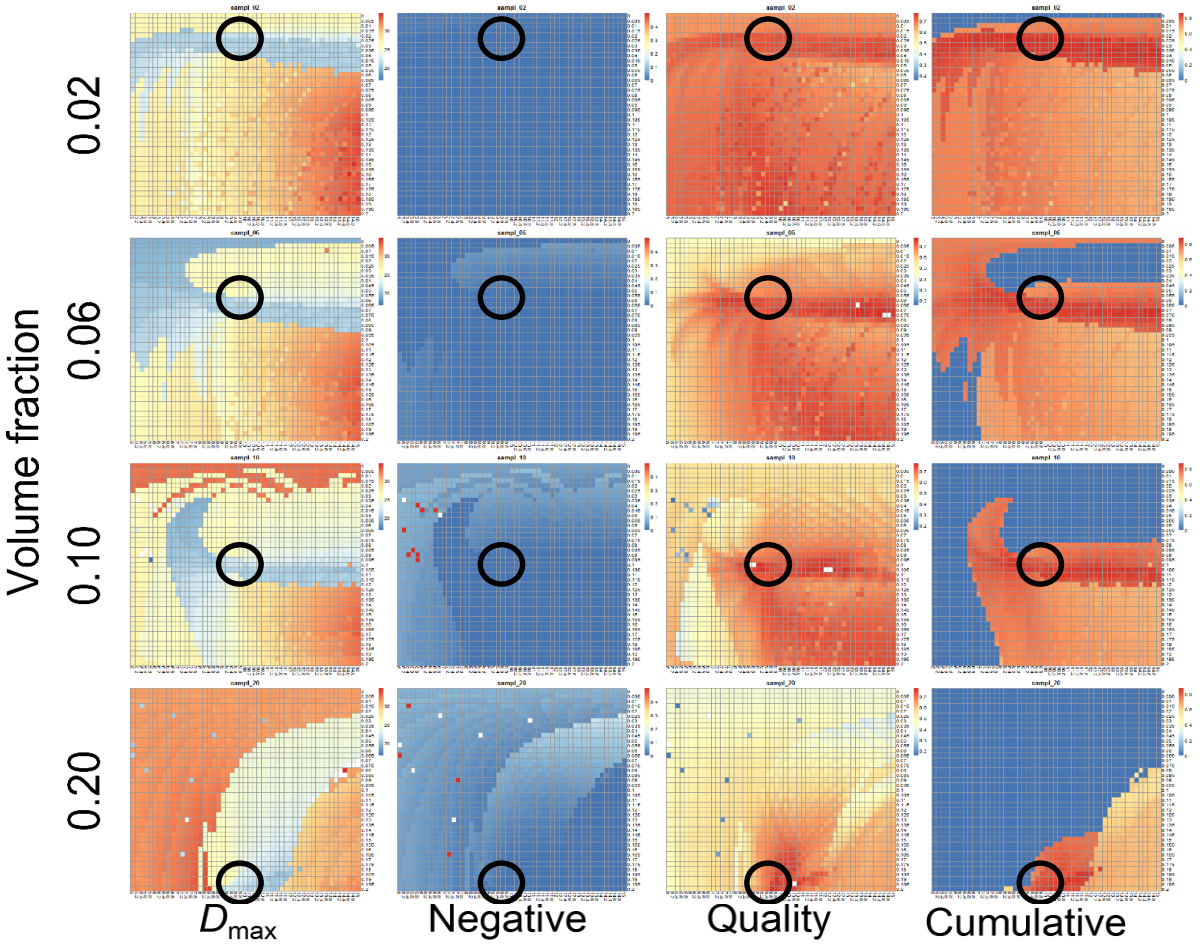

Fig. 12. Estimation of the solution parameters and cumulative quality estimator

3.3.3 Improvements of the algorithm

3.3.3.1 Modified Structure Factor

The structure factors approximation based on the Percus-Yevick approximation for hard spheres does not always provide sufficiently accurate description for experimental data measured from (non-spherical) macromolecules in solution. To overcome this issue we propose to use a Modified Structure Factor (MSF) that is, in essence, a structure factor adjustment aimed at minimizing the discrepancy between approximated and experimental structure factors. The idea of this approach is to modify the values of the given approximate structure factor in the local minima and maxima preserving the smoothness of the function (Fig. 13). The coefficients of this modification is also determined via simulated annealing and the smoothness of the function is assured by the spline interpolation. This procedure is computationally expensive and is therefore performed only during the last iterations of the entire process, when the changes in other parameters (e.g. R_{hs} or v) are small.

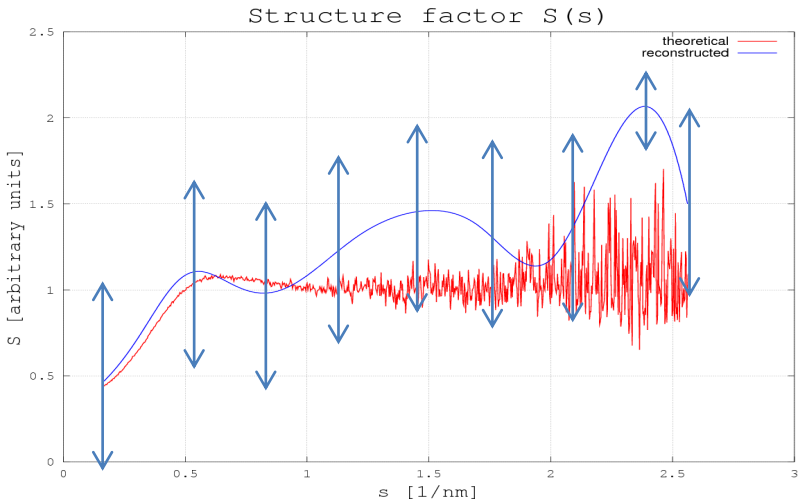

Fig. 13. Modified Structure Factor principle

3.3.3.2 Multiple scattering curves

In the course of a typical SAXS experiment several concentrations of a sample are measured. In terms of structure factor this means that we can proceed with several scattering curves having the same form factor but different structure factors. This idea has two implications: first, as more information is available the accuracy of form factor determination increases, and, second, we can use the assumption that among structure factors only the volume fraction is different, but the radius of interaction (R_{hs}) stays the same, according to the Perkus-Yevivk approximation [63]. By the calculation pairwise ratio of scattering curves, we eliminate the form factor $P(s)$ and can determine the points where the structure factors are the same and therefore reconstruct the R_{hs} (Fig. 14).

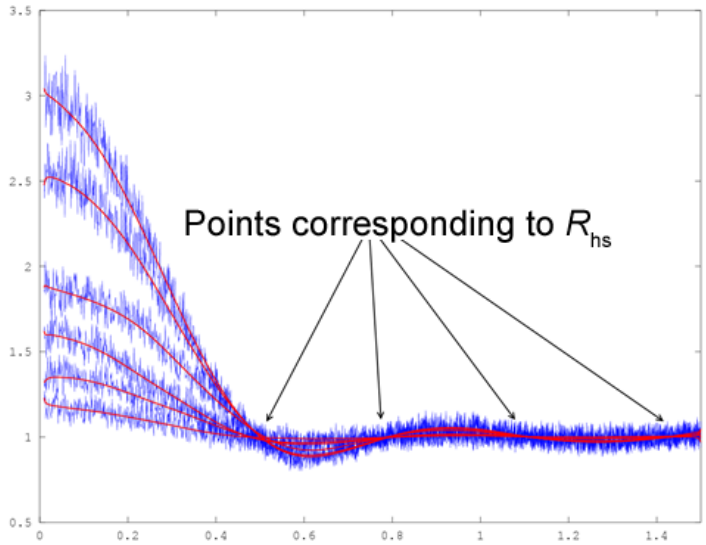

Fig. 14. Determination of R_{hs} with multiple curves

3.3.3.3 Optimization of the algorithm

The simulated annealing calculations may take very long time especially when using an iterative approach. To minimize the running time two most computationally expensive steps of algorithm were determined using Octave profiler:

i. calculation of the Modified Structure Factor (due to the use of the spline interpolation);
ii. approximation of structure factor (not very demanding by itself but repeated many times).

The first issue was solved by using MSF only during the last iterations and the second by pre-calculation of sine and cosine functions values used in estimation of the structure factor **S** (equation 12). Implementation of these measures decreased the calculation time from hours to minutes.

3.3.4. Synthetic and experimental tests

To check if algorithm works correctly a series of tests were performed with both synthetic and experimental data. The synthetic tests were conducted on a set of scattering curves from interacting spherical particles with radii of 5, 10 and 15 nm and volume fractions from 0 to 0.2. In all cases the algorithm was able to restore both the structure factor and distance distribution function. The example for the case of 10 nm spherical particles with volume fraction of 0.08 is shown in Fig. 15. The algorithm was able to correctly reconstruct distance distribution function with minimal discrepancy to the simulated data (Fig. 15A).

The operation of the algorithm on experimental data was tested on the scattering data from a high concentration (20 mg/ml) solution of bovine serum albumin (BSA) in HEPES pH 7.5 buffer and the results are shown Fig. 16. At this concentration the structure factor is already well pronounced and prevents the accurate determination of the distance distribution function with standard methods. Perkus-Yevik approximation cannot precisely describe the experimental structure factor, therefore in this case we applied Modified Structure Factor (Fig. 16B) approach as well as basic algorithm (Fig. 16A). Although the results obtained with basic algorithm are already much better than without the structure factor reconstruction, the solution has an overestimated D_{max} and negative values in $p(r)$ occur between 10 and 15 nm (Fig. 16A green curve). The MSF substantially improved the quality of solution and the final distance distribution function (Fig. 16B green curve) is almost the same as expected one from the BSA (Fig. 16B red

curve). The fit to the experimental data is presented in Fig. 16C and the experimental and reconstructed structure factors are compared in Fig. 16D.

3.4 Conclusions

The obtained results show that the proposed algorithm can quantify and reconstruct structure factor contributions and calculate distance distribution functions for both simulated and experimental data sets. The introduced improvements of the basic algorithm make it operation more stable and accurate. The application of the developed method to the experimental data allows deconvolution of the form and structure factor contributions to experimental scattering data. The program is planned to be made publically available upon its inclusion in ATSAS.

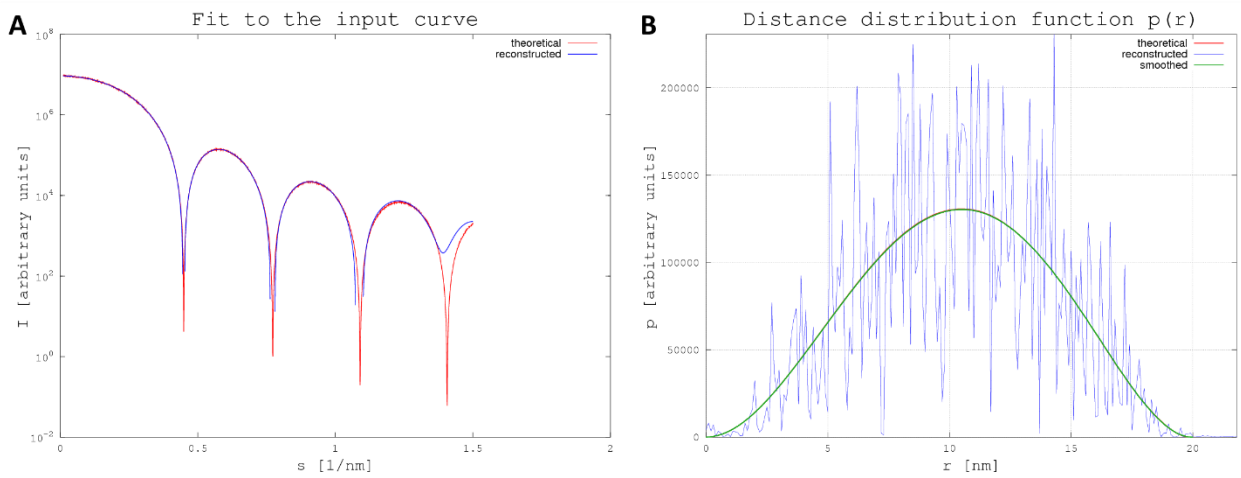

Fig. 15. Determination of distance distribution function of interacting spherical particles. A. Fit of the final solution (blue) to the test scattering curve (red). B. Reconstructed distance distribution function, blue – solution found by the algorithm, green – smoothed solution by GNOM, red – theoretical distance distribution function.

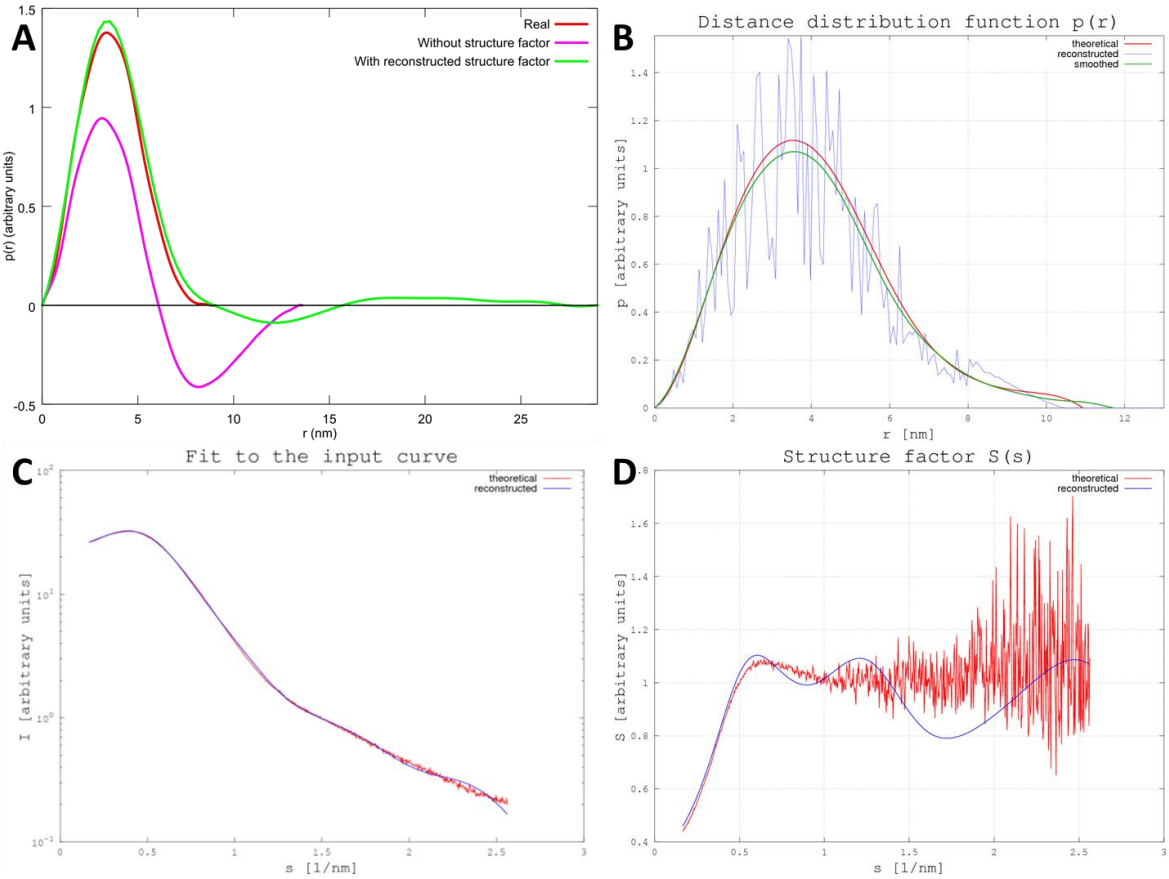

Fig. 16. The determination of distance distribution function and structure factor of high-concentration BSA. A. Without Modified Structure Factor (red curve, the actual distance distribution function free of structure factor effects;green the reconstructed p(r) using the basic algorithm; magenta p(r) with the influence of S(s)). B. With application of the modified structure factor (red curve – expected distance distribution function, blue – approximation found with the algorithm, green – smoothed solution by GNOM). C. Fit (blue) to experimental curve (red). D. Experimental (red) and reconstructed (blue) structure factors.

4. Chapter 4. Analysis of EOM capabilities

4.1. Introduction

Intrinsically disordered proteins (IDP) and other flexible macromolecular systems are objects of immediate interest in structural biology [64]. They perform many important functional roles in physiological processes from enzyme regulation to post-translational modifications [65], and are often implicated in severe diseases [66]. Investigation of the IDPs properties and their interactions with binding partners is challenging due to the limitations inherent in the standard high-resolution methods employed in structural biology. For example, molecular X-ray crystallography cannot be used to analyze IDPs, because crystallization is almost impossible. Even if crystals are obtained for an intrinsically flexible system, the crystal structure represents only one of the many conformations existing in the solution [67]. Presently, the most powerful and widely used techniques for structural characterization of IDPs are NMR spectroscopy and SAXS. Both methods have specific advantages and limitations and the modern approach is to combine such complimentary techniques and computational methods. Recently, an ensemble approach was developed to quantitatively analyze the SAXS data from flexible systems, first implemented in the ensemble optimization method (EOM) software [15, 16]. The aim of this project was to develop and conduct a series of test and case studies assisting in the design of a new enhanced version of EOM and also to assess the capabilities and limitations of the new program.

4.2. Structural characterization of intrinsically disorder proteins with small angle X-ray scattering

4.2.1 Intrinsically disorder proteins and their biological relevance

Historically, specific protein function was thought to be determined by a unique three-dimensional structure. This structure-function paradigm was supported by the determination of high-resolution 3D structures of thousands of proteins solved primarily by X-ray crystallography. However, as the number of observed structures increased it became apparent that not all proteins have rigid and well-defined structure across the entire sequence. On the contrary, many protein structures in the Protein Data Bank (PDB) contain missing regions of electron density. A typical explanation for missing density is the lack of coherent scattering caused by variations

in atomic position for regions of sequence across the different molecules in the crystal. In other words, missing electron density can be ascribed to the flexible or disordered nature of these fragments and their lack of the rigid 3D structure under physiological conditions [64]. According to recent bioinformatics studies more than 25% of eukaryotic proteins are disordered and more than half have long disordered fragments [68]. These unstructured regions have been termed natively unfolded [69], natively denatured [70] or intrinsically disordered [71]. In this work the latter term is used as it has become the most popular.

By definition, intrinsically disordered proteins (IDPs) or intrinsically disordered regions (IDRs) of folded proteins (e.g. loops or interdomain linkers) do not contain well-established secondary or tertiary structure. Instead they exist as an ensemble of conformations in solution. For these proteins the absence of well-defined 3D structure does not prevent them from performing their biological function, but instead is essential for the function. Many IDPs undergo disorder-to-order transitions upon binding to specific partners, among which are other proteins, nucleic acids and metals [64]. Also, posttranslational modifications frequently occur in disordered regions [65]. Obradovic and co-workers analyzed 98 IDPs and IDRs and assigned them to 28 functional classes [72], the largest being protein-protein and protein-DNA interactions and phosphorylation. In fact, many functions of disordered proteins, such as regulation, signaling and recognition, cannot be fulfilled with stable and rigid 3D conformations. Naturally these functions are very important and their disruption leads to various diseases, for example disordered protein p53 is associated with cancer and amyloid-β and τ proteins are linked with Alzheimer disease [66].

To sum up, the study of the properties and structural features of disordered proteins is important for a better understanding of many crucial processes in living organisms, as well as for the prevention and treatment of severe diseases. However the nature of such proteins limits their study with X-ray crystallography, the most common technique in structural biology. In the next section we consider biophysical techniques used for characterization of IDPs and their joint use with SAS.

4.2.2 Methods of structural characterization of IDPs

Combinations of computational and experimental techniques are often applied to the analysis of unfolded proteins including bioinformatics predictors, NMR spectroscopy, SAXS and Förster resonance energy transfer (FRET) [67].

The first step usually involves a bioinformatics assessment of the location of disordered fragments based on the protein amino acids sequence with several online tools such as PONDR-FIT [73]or IUPred [74]. Of course, these tools do not provide exact information on the presence and location of disordered regions, but they do help to see the general picture and to select the experimental methods for further research. Besides bioinformatics techniques the initial stage may include some common biophysical methods such as circular dichroism (CD) or infrared spectroscopy (IR). These methods yield information on the secondary structure and can identify low and high random coil content, providing an indicator of disorder. Information on the molecule dimensions, e.g. hydrodynamic radius, obtained with static light scattering (SLS) can also indicate disorder, because unfolded proteins tend to have much larger sizes than that of the folded state.

The most powerful and widely used high-resolution technique for structural characterization of IDPs is NMR spectroscopy [75]. The main disadvantage of the technique for structural analysis is the limitation on size of the protein to about 60 kDa. FRET is a single-molecule fluorescence technique that yields the distance between residues labeled with fluorescent dyes [76]. As fluorescence intensity increases with decreased distance, the proximity of labeled residues can be determined. Thus, similarly to SLS, FRET can be used for the estimation of the molecular dimensions, and due to the single-molecule nature this technique allows to obtain not only the average values but distribution of distances [77]. The disadvantage of FRET is the protein labeling with the dyes, which can be costly and difficult as well as can cause alteration of the protein properties.

Contrary to NMR spectroscopy, SAXS has no limitations for the size of proteins. However, it has lower resolution compared to NMR or molecular crystallography. Other advantages of SAXS are that measurements are performed in solution and without requiring a specific label (unlike FRET). A single SAXS experiment takes seconds to minutes for data acquisition, thanks to high brilliance of modern synchrotron radiation sources and significant improvements in SAXS beamline automation. This allows one to perform measurements in various conditions and for many protein isoforms or mutants in reasonably short times [14].

4.2.3 Application of SAS for the characterization of IDPs

Traditionally SAS is applied for the structural analysis of folded proteins and macromolecular complexes in monodisperse solutions, allowing the overall shape and low-resolution

structure to be determined. Under these conditions the particles in solution can be considered identical and a low-resolution model can be reconstructed from the scattering curve *ab initio* or through hybrid modelling using rigid bodies. The former method is used when there exists no *a priori* knowledge about the particle, and the latter when complex components or protein subunits are available. The application of neutron scattering with sample or solvent deuteration is also very useful for the reconstruction of multicomponent macromolecular complexes [19], but is not widely applied to IDPs. For that reason in this chapter we consider only SAXS data analysis approaches for unfolded proteins.

The traditional SAXS approaches mentioned above are not applicable to solutions of polydisperse particles, as in this case the scattering pattern is averaged over non-identical components. The structure of the individual particles in a polydisperse solution cannot be determined from the scattering curve alone, however, if the scattering profiles of the components are known it is possible to determine their volume fractions. In this case the scattering profile can be presented as a linear combination of the scattering of a system of components according to equation (21):

$$I(s) = \sum_{N}^{k=1} v_{\mathrm{k}}\, I_{\mathrm{k}}(s) \tag{21}$$

where N is the number of species in solution, v_k is the volume fraction of a single species and $I_k(s)$ is its normalized intensity. This approach is used to determine the volume fractions of components in oligomeric mixtures [78], and can in principle be employed for the characterization of more complex systems, including flexible molecules such as IDPs or multidomain proteins with flexible linkers. However, for IDPs, as each component k is a single conformation of the protein, and the number of conformations can be enormous, direct determination of the volume fractions is not feasible.

The scattering curve of a disordered protein is a sum of the patterns of the individual conformations present in solution. An example of such an average is shown in Fig. 17A, where scattering profiles of ten random conformations of a synthetic 100 amino acid-long polyalanine chain (black lines) extracted from a pool of 10000 conformations are shown alongside the averaged scattering profile of all structures in the pool [14]. While the scattering curves of individual conformations have distinct features, their average is smoother and almost featureless. Moreover, the differences between individual curves are observed across the entire momentum

transfer range, including the initial low-angle region of the profiles, corresponding to the largest intramolecular distances. Thus, while the scattering profiles of individual conformations of disordered proteins demonstrate diversity in size and shape, the averaged curve shows the averaged values. Although a smooth and featureless scattering curve is often a sign of a protein being in a disordered state, it is worth mentioning that this in itself is not sufficient for a clear distinction between folded and unfolded proteins.

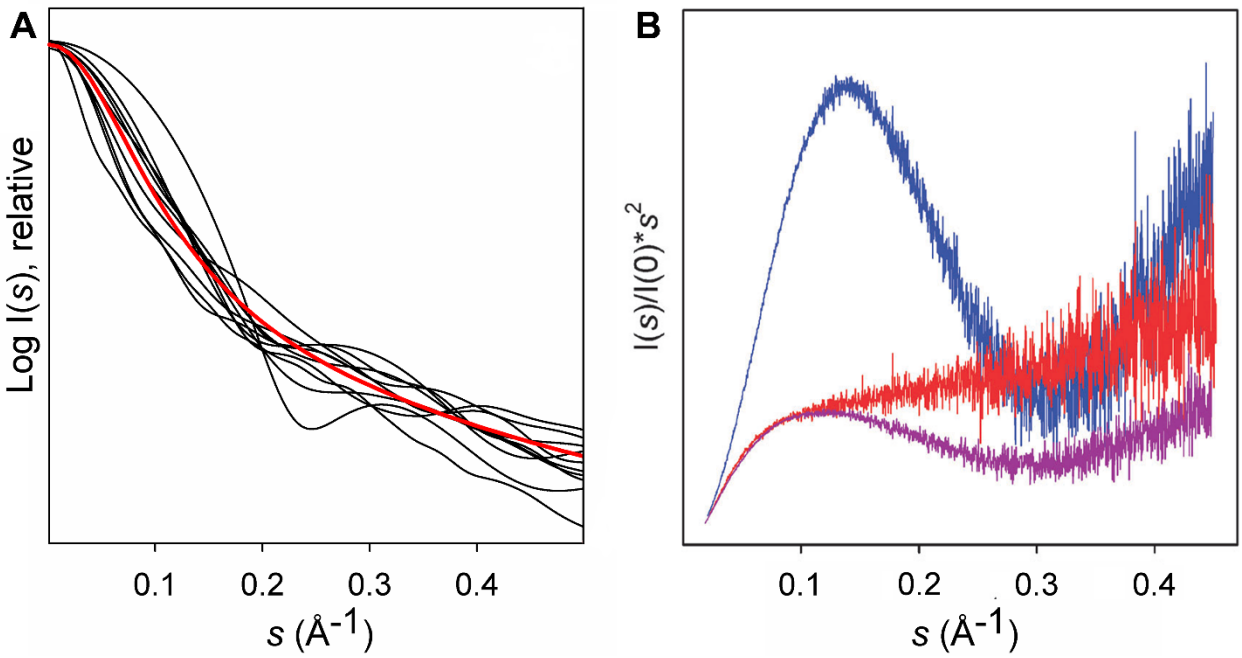

Fig. 17. A. Individual SAXS profiles (black) of ten randomly selected chains and averaged curves of 10,000 conformations (red) B. Kratky plot for three constructs of Src kinase. The globular SH3 domain (blue), the fully disordered unique domain (red), and a construct joining both domains (purple). The prototypical features of globular and disordered domains are combined in the partially folded construct. [14]

A more reliable way to determine from the scattering data whether a protein is disordered is to utilize a Kratky plot representation – the plot of the product of scattering intensity and squared momentum transfer ($I(s)\cdot s^2$) as a function of s [23]. This is commonly used for qualitative discrimination of folded and disordered states because the shape of the plot is defined by the degree of extent of the protein [79]. For globular proteins the scattering intensity decays at higher angles approximately as $1/s^4$ and therefore the corresponding Kratky plot has a distinct bell-like shape with a well-defined maximum. Conversely, the scattering curve of an ideal Gaussian chain displays an asymptotic behavior as $1/s^2$ and therefore has a plateau at large s values in the Kratky plot. Unfolded proteins are not ideal Gaussian chains and in the Kratky plot exhibit a plateau only at low values of momentum transfer followed by a monotonic increase.

In Fig. 17B typical experimental Kratky plots for globular (blue), partially unfolded (purple), and completely disordered (red) proteins are presented.

Yet another approach for characterization of unfolded proteins is the estimation of molecule compactness by the comparison of the experimental R_g of the sample with theoretical estimates for globular and disordered proteins. IDPs are expected to have larger geometrical parameters as they adopt more extended conformations. The most applied and convenient quantitative descriptor for compactness is R_g , which can be related to the number of residues in a polypeptide by Flory's equation [80]. This equation establishes power law dependence between the parameters:

$$R_g = R_0 N^{\nu} \tag{22}$$

where N is the number of residues in the chain, R_0 is a scaling constant, and ν is an exponential factor. The theoretically calculated value of ν for the Gaussian chain is 0.588 [81]. The measurements of R_g for 26 chemically denatured proteins with the length from 8 to 549 amino acids estimated ν to be 0.598 ± 0.028 and R_0 = 1.927 ± 0.27 [82]. The obtained values show that denatured proteins behave similarly to random coils in terms of R_g, but it is not clear whether the results for chemically denatured proteins can be applied to intrinsically disordered proteins. It has been established that chemical denaturation agents such as urea or guanidinium chloride interact with backbone and/or side chain atoms leading to changes in Ramachandran populations [83]. In an NMR study chemically denatured samples of ubiquitin were shown to have larger populations of extended conformations compared to that of IDPs [84]. The R_g values of chemically denatured and natively unfolded proteins have also been compared using an ensemble approach [79], revealing a 15% increase in extended conformations in the ensembles of denatured proteins as compared to IDPs. This study provided new estimates for the Flory's equation parameters for IDPs: ν = 0.522 ± 0.010 and R_0 = 2.54 ± 0.01.

The methods described above provide only qualitative analysis of IDPs. For the quantitative characterization of flexible structures with SAXS the most effective method is the ensemble approach used in this study and reviewed in the next section.

4.2.4 Ensemble approach in the description of IDPs

The major challenge for a quantitative analysis of SAXS data for flexible systems lies in the treatment of a countless number of coexisting conformations. This issue cannot be solved

with traditional methods of SAXS data analysis including Kratky plots, which yield only qualitative assessment of the sample. The best approach to the quantitative characterisation of disordered proteins is to describe the system as an *ensemble of conformations* [15, 79, 85, 86], also called the *supertertiary structure* [87]. The methods based on this idea were introduced quite recently and were initially applied in NMR studies [85]. The first implementation of an ensemble approach for SAXS was the Ensemble Optimization Method (EOM) introduced in 2007 [15]. Since then several other methods of ensemble-based representation of disordered structures have been developed [88-91], making this approach one of the most popular for the analysis of IDPs. The usage of the experimental data obtained by complimentary methods (NMR, SAXS, circular dichroism (CD), bioinformatics methods, etc.) allows for reliable reconstruction and validation of the properties of the modelled structures.

In SAXS data analysis, the ensemble description is used to represent a polydisperse mixture of conformations, with each conformer yielding a single scattering pattern. Thus equation (21) can be applied to an ensemble of N components each with its own scattering intensity. The problem is reduced to the determination of the volume fraction of each conformation in solution (i.e. of the ensemble components). Based on this idea several computational methods for the SAXS data analysis of flexible molecules have been developed, following the strategy consisting of three major steps:

1. Generation of a large pool of diverse structures approximating the conformational space of the studied polypeptide;
2. Calculation of the scattering properties of each individual conformation;
3. Selection of an ensemble of structures that fits the experimental data using an optimization method.

Each step of this algorithm can be performed in different ways, so further we describe the existing implementations of this strategy and their distinctive features. As the EOM was first implementation of ensemble approach used in SAXS data analysis and assessment of its capabilities and limitations are the aim of this project, it is presented in the dedicated section.

Among the other methods, the main feature of *minimal ensemble search (MES)* is prevention of overfitting of the experimental data [88]. The Bayesian-based Monte Carlo algorithm employed in the *basic-set supported SAXS (BSS-SAXS)* estimates not only the fraction of each conformation, but also their uncertainties [89]. In the *ensemble refinement of SAXS (EROS)* the

initial sampling is performed with coarse-grind approach and the obtained structures are clustered to determine independent states and the optimization to fit experimental data is done by varying the relative population of the clusters [90]. The more complex algorithm of pool generation is implemented in *ENSEMBLE* method: 5000 structures for the pool selected from the 100000 "initial soup" [91]. In the course of optimization process the pool is updated by removing structures that are not used in fitting and adding conformers similar to conformers in selected ensemble. The program *Flexible-Meccano* (FM) samples conformational space by assembling repeatedly rigid peptide units according to the Ramachandran's plot and avoiding the clashes by a coarse-grained description of the side-chains.

4.2.5 Ensemble optimization method (EOM)

EOM was the first method developed to analyze SAXS data of flexible structures using the ensemble approach [15]. The basic idea of this method is the determination of the properties of the ensemble of coexisting conformations, whose combined scattering profile fits the experimental SAXS curve.

The scattering profile of the ensemble is calculated using the individual scattering patterns and assuming equal populations of each conformer:

$$I_{EOM(s)} = \frac{1}{N}\sum_{n=1}^{N} I_n(s) \tag{23}$$

where $I_n(s)$ is the scattering from the *n*-th conformer. The number of conformations in the ensemble (N) is determined during the optimization procedure. The conformations are selected from a large pool, which represents the maximum possible flexibility based on the topology of the protein. In addition to the topology additional *a priori* information about the system may be utilized when available (e.g. distances derived from FRET or NMR), and in such a case EOM is no-longer a fully model-independent approach. The scattering pattern is pre-computed for each structure in the pool and used in the optimization procedure.

The optimization is performed through the genetic algorithm (GA) and at each optimization step a potential solution is examined by EOM as an ensemble of N different structures/conformations. The structures can be either conformers of the same molecule from one pool or a mixture of oligomers from several independently constructed pools. The optimization strategy consists of minimization of χ^2 discrepancy according to:

$$\chi^2 = \frac{1}{K-1}\sum_{i=1}^{K}\left[\frac{\mu I(s_i) - I_{\exp}(s_i)}{\sigma(s_i)}\right]^2 \quad (24)$$

where K is the number of points in the scattering curve, μ is a scaling factor, $I(s_i)$ is the calculated scattering from the ensemble, $I_{exp}(s_i)$ is the experimental scattering and $\sigma(s_i)$ are experimental errors. Therefore the ensemble selected in the optimization process is a subset of the starting pool(s), which best fits the experimental data. However, it should be stressed that the ensemble members themselves serve only as a tool to construct distributions of structural parameters, which provide the basis for further analysis. The reported parameters consist of radius of gyration R_g and maximum dimension of the particle D_{max}. The distributions of R_g and D_{max} of the selected ensemble are compared with those of the initial (random) pool and provide estimations of the particle compactness and flexibility [15].

Although the original EOM has been widely applied in the analysis of SAXS data for disordered and flexible macromolecular systems [92, 93], practical applications also revealed some limitations of the method. To overcome these limitations a new version EOM 2.0 was developed [16]. In this implementation several important improvements were introduced including a new algorithm for the generation of missing fragments (in two modes – random and native, which are defined by the distribution of dihedral angles in the models), generation of symmetric and asymmetric oligomers, optimization of the ensemble size, usage of multiple pools in the optimization process and quantitative numerical estimators of the sample's flexibility and compactness. The aim of the project presented in this chapter was to develop tests for the new version of EOM and determine its capabilities and limitations.

4.3. Determination of the EOM capabilities

Since its introduction in 2007, EOM is being actively used by the structural biology community and the original publication was cited more than 400 times according to Google Scholar, for example in papers [94-96]. However, feedback from the community suggested some possible improvements and EOM 2.0, with several new features, was released as a part of ATSAS 2.6 package. To make sure that this implementation is suitable for the analysis of flexible macromolecules it must satisfy two main conditions common for any ensemble-based method: the initial search pool must adequately describe the conformational space and correctly select the ensemble that is derived appropriately form the experimental data. To check if those

prerequisites are met by EOM 2.0 the set of tests was performed and these were analyzed with the statistical software R [97].

4.3.1 Accuracy of unfolded proteins conformational space sampling

It is important that the generated structures in the pool correctly represent the properties of unfolded proteins. For IDPs the population of end-to-end distances is expected to follow that of a Gaussian distribution, with a mean squared end-to-end distance equal to 70 (± 15) $Å^2$ × N, where N is the number of amino acids in the polypeptide chain [98]. To check whether this condition is met two pools of 10000 polyalanine chains of sequence length 100 or 500 residues were generated in random mode and for each sequence length the procedure was repeated 5 times. The distributions of the end-to-end distances for the EOM pool (blue histogram) and Gaussian chains (red curve) are shown in Fig. 18.The quantitative assessment of the discrepancy between obtained (EOM) and theoretical (Gaussian) populations were calculated as RMSDs between corresponding distributions and they were equal to 1.6×10^{-4} (for 100AA chain) and 1.2×10^{-4} (for 500AA chain) respectively. The mean squared end-to-end distances were 5590 $Å^2$ (~56 × *N*, *N*=100) and 33060 $Å^2$ (~66 × *N*, *N*=500) for smaller and larger chains, respectively, in agreement with the experimental data for unfolded proteins [99]. This result indicates that a pool of 10000 conformations is sufficient to approximate the conformational space of unfolded proteins. A positive skew of the EOM end-to-end distance distribution relative to the normal distribution is caused by the fact that EOM builds models according to the distribution of dihedral angles specific to IDPs and avoids steric clashes in the generated structures.

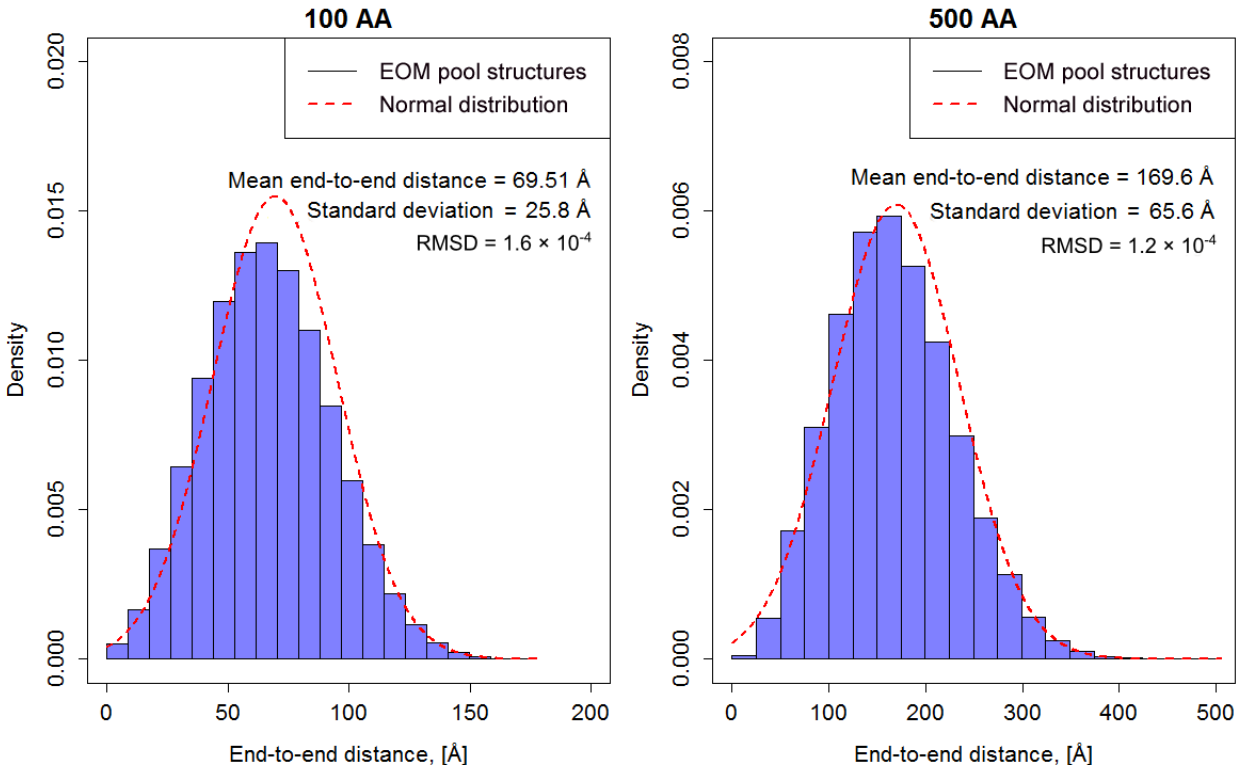

Fig. 18. Comparison of the end-to-end distances distributions for EOM pool and theoretical distributions of Gaussian chains for 100 and 500 amino acid length chains

4.3.2 Evaluation of ratio between number of amino acids and R_g of EOM generated structures

The radius of gyration, R_g is one of the most important parameters for the analysis of macromolecules by SAXS. According to Flory's equation (22) the R_g of a polypeptide chain has a power law dependence on the number of amino acids [80]. To investigate the concordance of EOM 2.0 generated structures with this law, 12 pools of 10000 conformers of polyalanine were constructed for sequences of varying length (10, 20, 50, 100, 200, 500 amino acids) in both random and native modes. The results of the modelling is shown in a log-log plot in Fig. 19 where the average R_g of the pools generated in random (blue curve) and native (green curve) modes are located between theoretical estimations for random coil (purple) and globular proteins (red). For the determination of the average R_g of globular proteins the coefficients from Flory's equation for globular proteins were used (ν = 0.34 and R_0 = 2.83 [100]). Besides the average values, the upper and lower quartiles of the pools distributions are demonstrated as triangle marks in the plot. The quartiles were chosen instead the extreme values to eliminate the outliers.

The average R_g of the pools is found to be in the agreement with the theoretical estimations established by Flory's law. Thus, the search pools generated in by EOM 2.0 do represent the properties of disordered proteins. Moreover, the R_g distributions of the pools spread wide enough to include partially folded proteins.

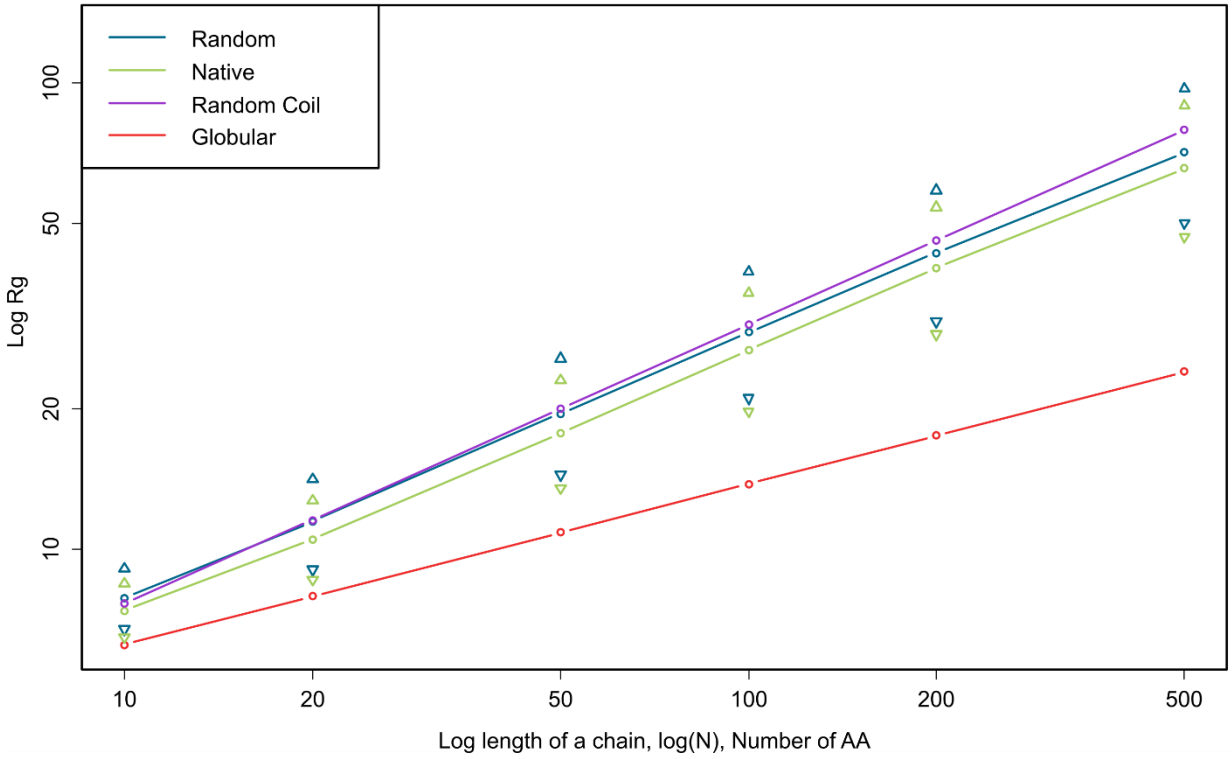

Fig. 19. Relationships between R_g and length of polypeptide chain (log-log plot). Curves corresponding to the EOM pools are shown in blue (random mode) and green (native mode), triangular marks corresponds to the upper and lower quartiles of the pool R_g distributions. Theoretical estimations for globular proteins (red) and random coil (purple).

4.3.3 Discrimination between distinct conformations in a mixture

Due to the ability of EOM to incorporate models derived from high-resolution methods (for example domains or fragments of proteins) into the optimization procedure the method can be used for distinguishing between different conformations of folded proteins. One of the possible applications of this approach is to determine the proportion of open and closed conformations of a protein in a given mixture. As an example we used the calcium binding protein calmodulin, using the high-resolution structures of the 'open' (1CLL R_g = 16.7 Å) and 'closed' states (1CTR,R_g = 22.6 Å). The scattering curves from the two states were computed using CRYSOL [29] and averaged with PRIMUS [78] to simulate mixtures of open and closed con-

formations. Calmodulin is an excellent case to test the capacity of EOM to extract distinct conformational states because the protein has two well defined EF-hand domains connected by a labile helical linker. To determine, which parameters are optimal we created three pools of 30000 structures each with different flexible segment lengths between the two EF-hand domains (0, 2 or 10 amino acids). Then the curves corresponding to opened and closed structures and a mixture of them were used as input for EOM. The obtained distributions of R_g presented in Fig. 20 show that in all cases EOM was able to resolve two conformations both in pure and mixture states. The constructs with longer flexible segments have better fit (lower chi-value) and the peak of the R_g distribution were closer to the CRYSOL-calculated R_g.

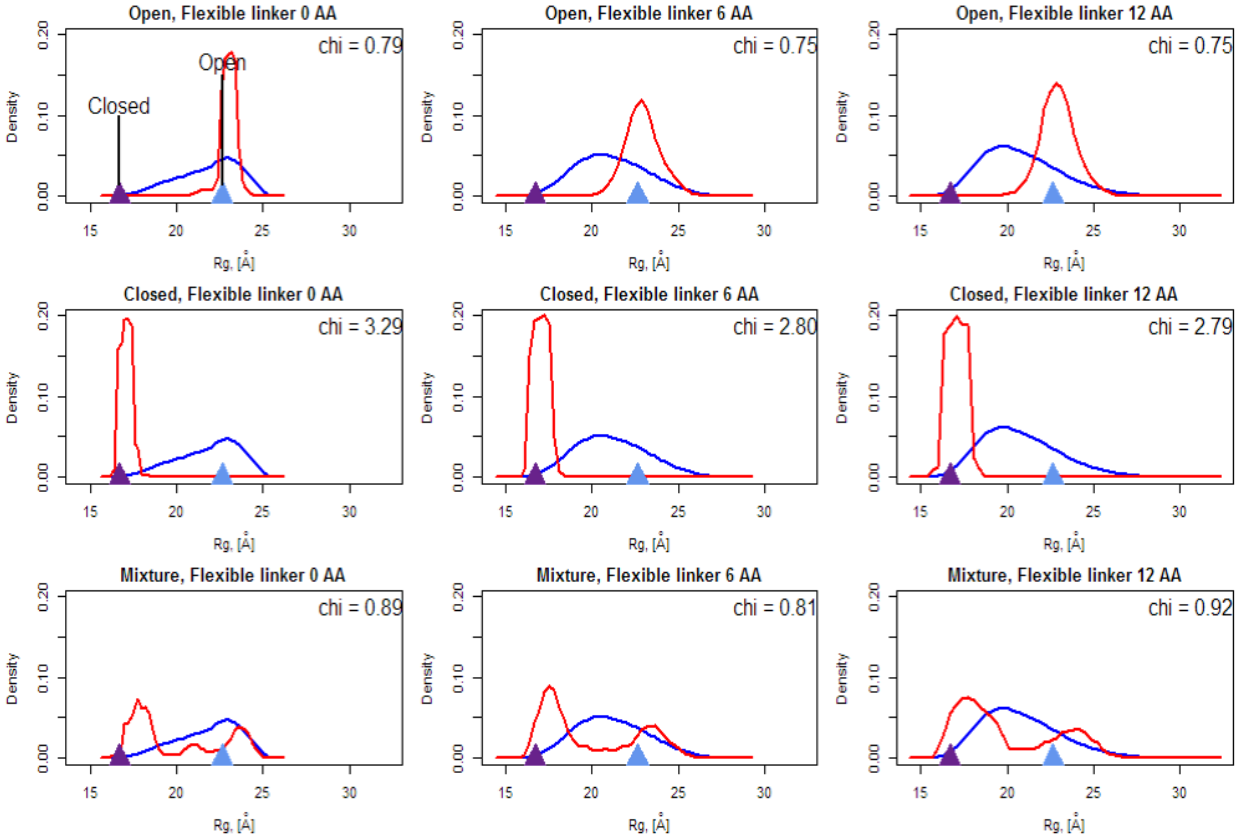

Fig. 20. Resolution of open and closed conformations of calmodulin. Blue curves are R_g distributions of used pools, red curves are R_g distributions of the selected ensembles, purple and light blue triangle marks are R_g of open and closed conformations respectively.

4.3.4 EOM resolution

The ability of EOM to resolve different structures in a mixture has already been successfully utilized [16], but the question remains as to what is the true resolution capability of EOM? In other words, what is the minimum difference in R_g between two subpopulations in a mixture

that can be distinguished? To test the resolution of EOM 2.0, an initial pool containing 10000 models of polyalanine (for 100 and 500 amino acid long chains) was generated and a subset of conformers representing two subpopulations, each with a different mean R_g and standard deviation, extracted. The theoretical scattering intensities of the members of the subset were calculated by CRYSOL and averaged, producing a simulated test data set. This data was then used as input for the EOM selection from another independently generated pool, and the resulting R_g distributions of the selected ensembles were examined. The test was repeated several times varying the difference between the mean R_g values of each subpopulation and their standard deviations. The R_g distributions produced from these tests (Fig. 21) demonstrate that the bimodal distributions expected are indeed observed, indicating that subpopulations of structures can be resolved. The results show that the resolution does not depend on the width (standard deviation) of the subpopulations, unless they intersect, but strongly depends on the absolute difference in the values of mean R_g. Two sub-populations should have a relative difference greater than ~2 times the standard deviation of the pool from which they come from in order to be distinguished. This result did not depend on the size (number of amino acids) in the polypeptide chain.

4.3.5 Robustness of the method and impact of noise

For any method that is based on fitting experimental data with generated models, an assessment of the effects of experimental signal-to-noise ratio is essential. To test the robustness of EOM we conducted the following simulations. First, a pool of 10000 structures for 100 amino acid long protein was generated, and then a structure with an R_g close to mean for the pool (33.1 Å) was selected. The scattering curve of the structure was modified by cutting lowest angles up to 0.01 Å and adding noise from 0 to 5% of the value for each point in order to simulate errors in experimental data (Fig. 22A). For each noise level these manipulations were repeated 50 times and the obtained curves were used as an input for EOM. In this case we used the same initial pool for the EOM genetic algorithm, so in ideal case the program should select the same curve and structure that was used as input, irrespective of the noise level. In order to check this we compared the mean R_g of the selected ensemble with the initial structure and calculated the relative error. A box plot visualizing the obtained results is shown in Fig. 22A and we can conclude that with zero noise level EOM indeed identifies the initial structure. Even at a high noise level (up to 20%) the average R_g for the final ensemble solution is found to be in a very good

agreement with the R_g computed from the Guinier region. It can be concluded that EOM 2.0 is able to provide reliable solutions up to a 20% noise level in the experimental data.

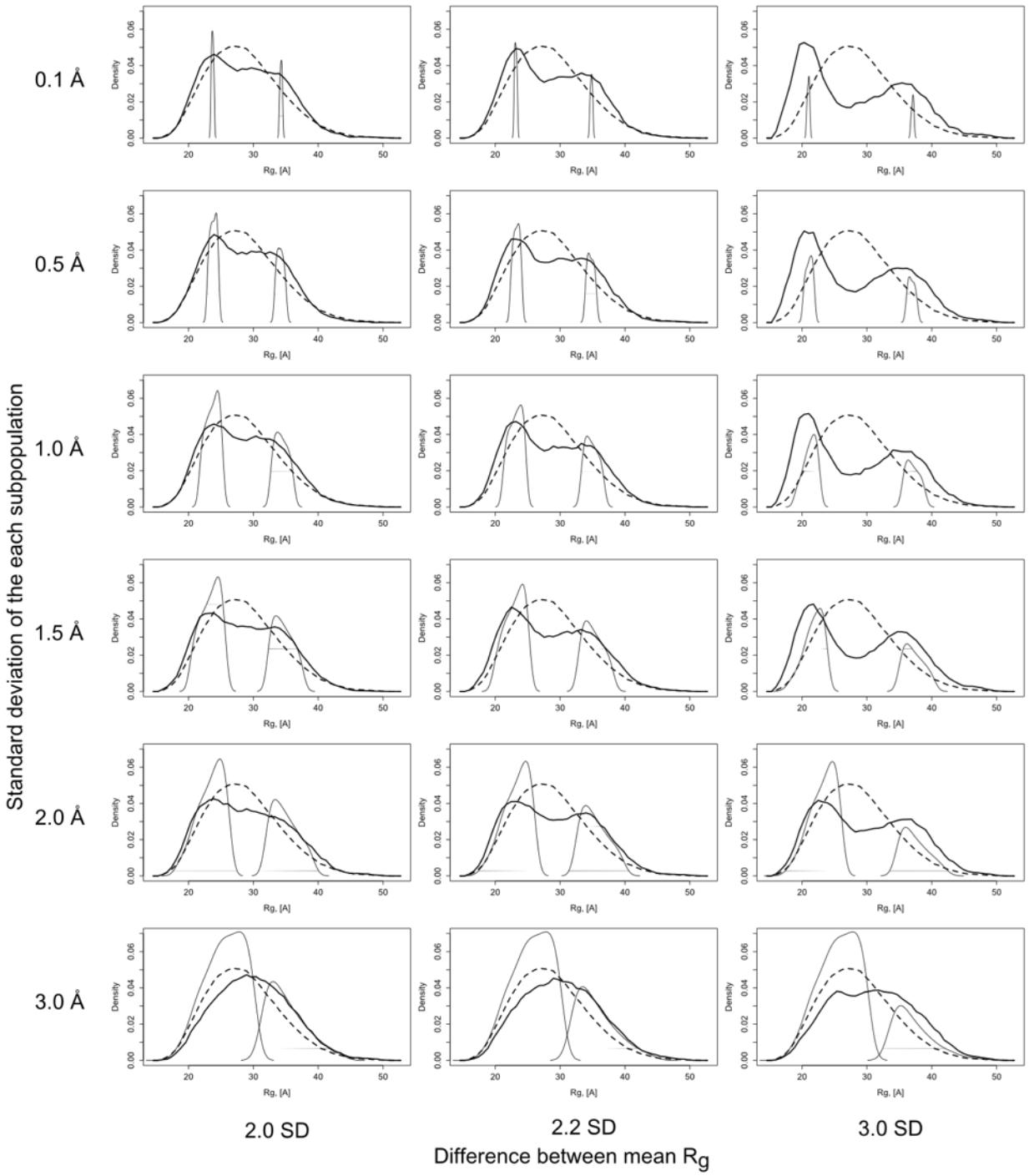

*Fig. 21. Distribution of the pools (*black dashed lines*) and selected ensembles (*black solid lines*) with various standard deviations differences between mean* R_g *of the subpopulations (*grey solid lines*). The comparison shows that the EOM 2.0 resolution depends on the absolute difference between their mean* R_g*, but not on the width (standard deviation) of subpopulations, unless they intersect.*

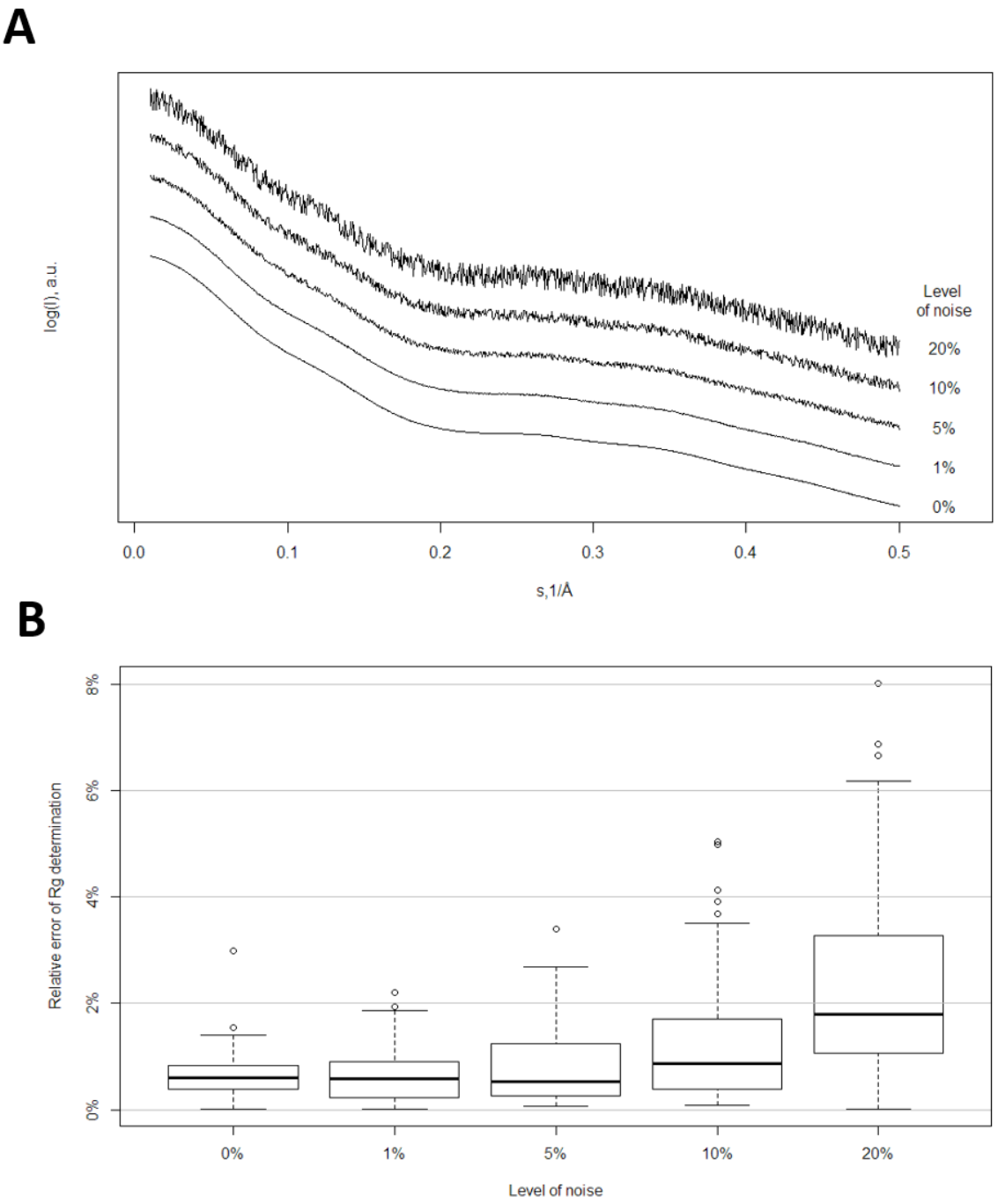

Fig. 22. (A) Comparison of the scattering curves used to check the robustness to noise of EOM 2.0 in the case of complete absence of noise (0%) and with 1%, 5%, 10% and 20% random noise respectively. (B) Dependence of relative error in the R_g *determination on level of noise.*

4.4. Conclusion

Structural characterization of disordered proteins and of other flexible structures is an important topic in modern structural biology. SAXS is one of the most powerful techniques for analysis of such molecules and the ensemble approach is an optimal way to quantitatively characterize their properties in solution. The first implementation of an ensemble approach for SAXS was the Ensemble Optimization Method (EOM), recently updated to a new version, EOM 2.0. The tests presented in this chapter have shown that EOM 2.0 is able to correctly represent the properties of the unfolded proteins, resolve distinct conformations and subpopulations of flexible structures. The method was shown to be robust to the noise in scattering curves up to 20% of the intensity value. The results of these tests are included in the publication

Tria, Giancarlo, Haydyn DT Mertens, Michael Kachala, and Dmitri I. Svergun. "Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering." IUCrJ 2, no. 2 (2015): 207-217.

Chapter 5. Overview of collaborative projects

5.1. Introduction

Small-angle scattering is widely applied for characterization of structural properties of biological macromolecules and in the course of this PhD project several collaborative projects were performed with users of EMBL SAXS beamlines. The systems in this projects had very diverse nature including a globular fusion protein, an IDP and a macromolecular complex.

The experimental scattering data for all projects were collected at EMBL P12 SAXS beamline at storage ring PETRA III (DESY, Hamburg, Germany) with a PILATUS 2M pixel detector (DECTRIS, Baden, Switzerland) [101]. The measurements were performed with a sample-detector distance of 3.1 m and wavelength of 1.24 Å. A range of momentum transfer of $0.0028 < s < 0.45$ Å^{-1} ($s = 4\pi \sin(\theta)/\lambda$, where 2θ is the scattering angle) was covered with this set-up. Typically a sample was measured at least 3 different concentrations ranging from approximately 1.0 to 5.0 or 10.0 mg/ml. If the concentration dependency was observed the scattering data was extrapolated to zero concentration. To test for radiation damage, a one-second exposure time was divided into twenty 50-ms frames. In all cases no radiation damage was detected when the frames were compared. Basic data processing including normalization to the transmitted beam, radial averaging, subtraction of the buffer scattering, and scaling for sample concentration was done by automated data processing pipeline. Then the collected data was processed with PRIMUS [78] and the forward scattering $I(0)$ and radius of gyration (R_g) were calculated using the Guinier approximation [2]. The molecular mass of the samples was determined by comparing forward scattering with that of the reference protein (BSA) with known molecular mass of 72 kDa. The pair-distance distribution function $p(r)$ and the maximum particle dimension (D_{max}) were computed with the program GNOM [13]. The further steps of data analysis depended on the project and are described in the corresponding sections.

5.2. I27-PimA fusion protein

This project is presented in publication: *David Giganti, Jorge Alegre-Cebollada, Saioa Urresti, David Albesa-Jové, Ane Rodrigo-Unzueta, Natalia Comino, Michael Kachala et al. "Conformational plasticity of the essential membrane-associated mannosyltransferase PimA from mycobacteria." Journal of Biological Chemistry 288, no. 41 (2013): 29797-29808.*

Phosphatidyl-myo-inositol mannosyltransferase A (PimA) is an essential membrane-associated glycosyltransferase that initiates the biosynthetic pathway of key glycolipids/lipoglycans of the mycobacterial cell envelope [102], which are important molecules implicated in host-pathogen interactions [103]. This study was dedicated to a detailed investigation of the structural properties of PimA by single molecule force spectroscopy and SAXS. The experimental data for PimA in *apo*-form and PimA GDP complex was obtained at the EMBL X33 SAXS beamline at storage ring DORIS III (DESY, Hamburg, Germany) with a PILATUS 1M pixel detector (DECTRIS, Baden, Switzerland) [104, 105], and for the fusion protein I27-PimA at P12. The scattering curves and corresponding $p(r)$ functions are shown in Fig. 23 and determined parameters of the samples are presented in the Table 6.

Table 6. Parameters of PimA protein in apo-form, PimA GDP complex and I27-PimA fusion protein calculated from SAXS data

	PimA, apo	PimA GDP complex	I27-PimA
R_g nm	2.85	2.75	3.79
D_{max}, nm	10.5	10.0	12.5
Molecular mass, kDa	45	50	57

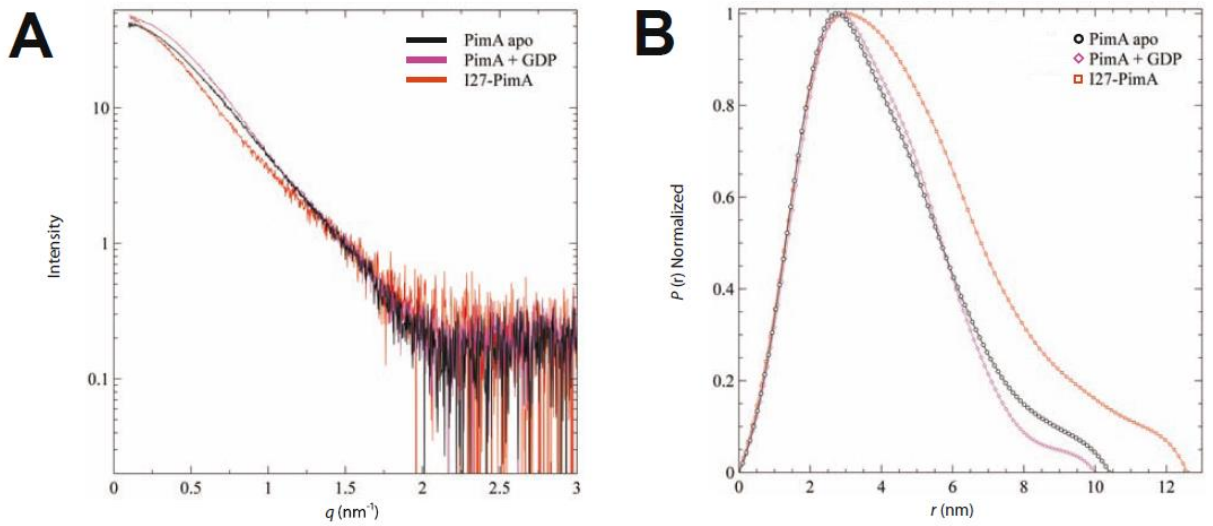

Fig. 23. Scattering data of PimA apo, PimA-GDP complex and I27-PimA fusion protein. A. Scattering curves of PimA apo, PimA-GDP and I27-PimA. B. P(r) function distributions of PimA apo, the PimA-GDP and I27-PimA.

A multiphase *ab initio* modelling was performed with MONSA program [27] on fusion protein I27-PimA data in order to elucidate the relative location of the subunits and orientation

of PimA. MONSA is based on DAMMIN software [27] and allows bead modeling of macromolecular complexes by simultaneously fitting multiple scattering curves. The search volume was defined as a sphere with the diameter equal to the maximum dimension of I27- PimA (12 nm) as computed by GNOM [13]. The search volume is filled with densely packed spheres of 1 Å radius. Each sphere within the search volume could correspond either to solvent, I27 or PimA. The minimization procedure fits the two scattering patters, PimA and I27-PimA to the corresponding calculated data obtained from computed structures by simulated annealing. The applied approach allowed us to determine the relative location of I27 and PimA within the fusion protein and orientation of PimA domains in the solution structures obtained (Fig. 24). The SAXS data and models are available in SASBDB under code SASDAS4.

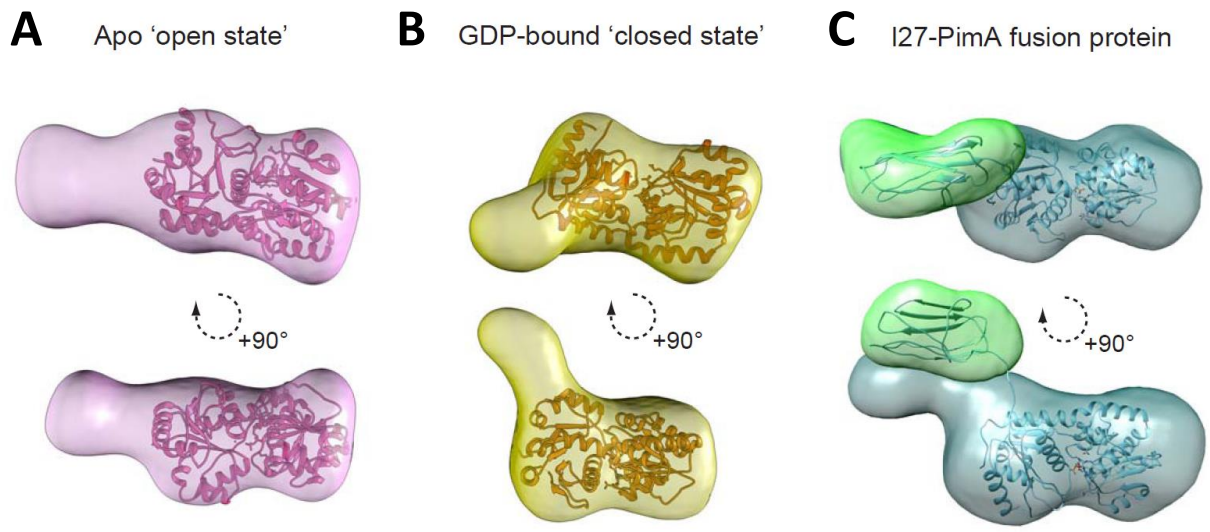

Fig. 24. SAXS based models of PimA in solution. A. Average low-resolution structure of PimA apo with the high-resolution crystal structure of PimA-GDP complex (PDB code: 2GEK) fitted by rigid body docking. B. Average low-resolution structure of PimA-GDP complex with the high-resolution crystal structure of PimA-GDP complex fitted by rigid body docking. C. Average low-resolution structure of I27-PimA fusion polyprotein with the high resolution crystal structures of I27 and PimA-GDP complex fitted by rigid body docking.

5.3. CD44 MEM-85 antigen-antibody complex

This project is presented in the manuscript "*Molecular mechanism for the action of the anti-CD44 monoclonal antibody MEM-85*" by Jana Skerlova, Vlastimil Kral, Michael Kachala, Milan Fabry, Ladislav Bumba, D. Svergun, Zdenek Tosner, Vaclav Veverka, and Pavlina Rezacova submitted to Biophysical Journal.

The hyaluronate receptor CD44 plays an important role in cell adhesion and migration and is involved in tumor metastasis. Binding of hyaluronate to hyaluronate-binding domain (HABD) of CD44 induces an allosteric conformational change, which results in CD44 shedding. Murine monoclonal antibody MEM-85 recognize epitope in human CD44 HABD and cross-blocks hyaluronate binding to CD44 also inducing CD44 shedding. MEM-85 has therapeutic potential, because in mouse model it inhibits growth of lung cancer cells. In this collaborative study, location of the epitope was suggested from the chemical shift perturbation NMR and mutational analysis. The aim of SAXS analysis was to further check the location of the epitope and relative positions of the CD44 HABD and MEM-85 in the complex using rigid body modelling.

The experimental scattering curve is shown in Fig. 25A, and the overall parameters of the complex are presented in Table 7. The low resolution *ab initio* model of the CD44 HABD – scFv MEM-85 complex was obtained by generation of 10 dummy atom structures using DAMMIN [27] and subsequent computation of an average model (using DAMAVER [106]) with a fixed core for another DAMMIN run. The final *ab initio* model of the complex has an elongated shape (Fig. 25C). To investigate the binding of scFv MEM-85 and CD44 HABD, rigid body modeling of the CD44 complex was performed with SASREF [31] using the NMR solution structure of CD44 HABD (PDB code 2I83[107]) without 15 flexible tail residues and a model of scFv MEM-85 created using the RosettaAntibody modeling server [108]. In the first modeling series, six options for paratope (heavy chain loops L1, L2, L3, H1, H2, and H3) were specified as possible sites of interaction with the epitope (residues 160-163 of CD44 HABD). In ten modeling runs, only two heavy chain loops (H2 and H3) were selected for the interaction with epitope. The next step of the modeling was performed with only these paratopes specified in the contact conditions and the models obtained in five runs were almost identical to each other. The final model of CD44 HABD and scFv MEM-85 was superimposed with the ab initio envelope using SUPCOMB [109] (Fig. 25C).

Table 7. Overall results of the SAXS experiment for the CD44 HABD – scFv MEM-85 complex

Structural parameters	
$I(0)$ (relative) from P(r)	8363 ± 135
R_g [Å] from P(r)	28.1 ± 0.5

$I(0)$ [relative] from Guinier	8235 ± 23
R_g [Å] from Guinier	26.9 ± 1.4
D_{max} [Å]	94 ± 10
Porod volume estimate [Å^3]	57000 ± 200
Molecular mass determination	
Molecular mass *MM* [kDa] from Porod volume (V_p·0.6)	34.2
Molecular mass *MM* [kDa] from forward scattering	43.8
Calculated monomeric *MM* from sequence	46.4

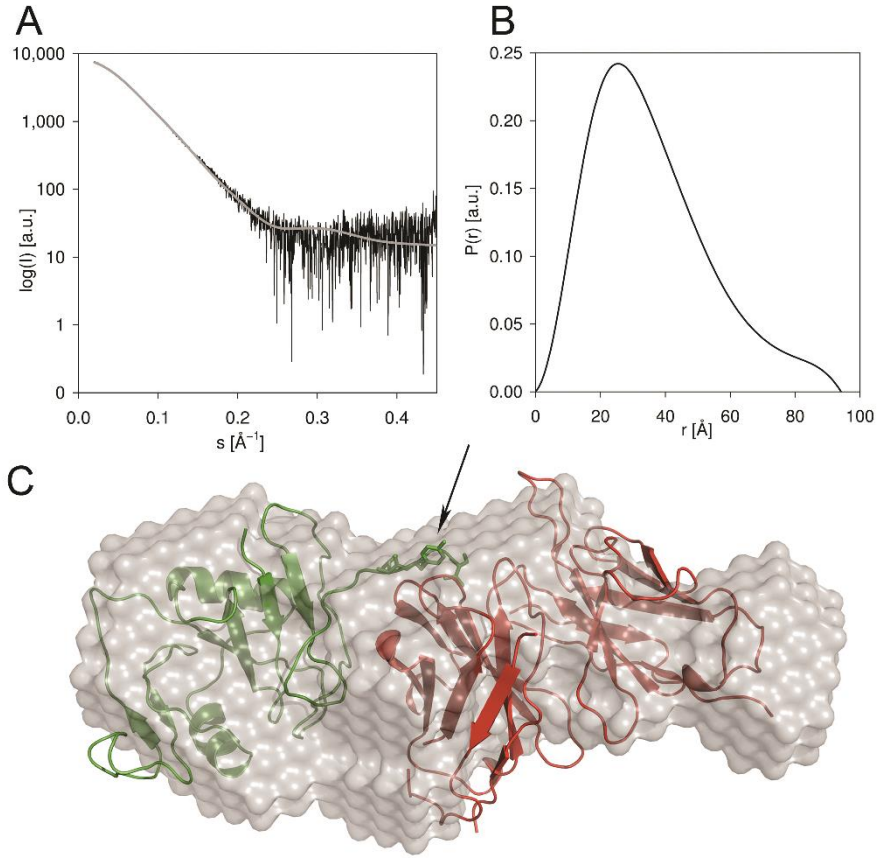

Fig. 25. SAXS data and rigid body model of CD44 HABD – scFv MEM-85 complex. A: The solution scattering pattern for the CD44 HABD – scFv MEM-85 complex (black) is shown with the fit of the theoretical scattering of the rigid body model of the complex (shown in panel C) to the SAXS experimental data (grey), where χ^2 = 1.24. B: The plot of the pair-distance distribution function p(r) is shown for the CD44 HABD – scFv MEM-85 complex with a maximum particle distance (D_{max}) of 94 Å. C: The rigid body model of the complex of CD44 HABD (green; PDB code 2I83[107]) and scFv MEM-85 model[108] (red) is shown fitted into the SAXS ab initio envelope. The disordered C-terminal portion of CD44 HABD (residues 164-178) is excluded from the model. The epitope is indicated with an arrow, and residues Glu160, Tyr161, and Thr163 are shown as sticks.

The scattering curve calculated from the rigid body model of the complex in which the epitope was defined as residues 160-163 of CD44 HABD exhibited a very good fit (χ2 = 1.24;

see Fig. 25C) and was superimposed well with the ab initio SAXS envelope. When the modeling was performed with the epitope defined as residues 38-42 of CD44 HABD, which had been previously identified as important for MEM-85 binding [110], the modeling program SASREF [31] could not fit the experimental scattering data with any model with a reasonably good fit even after 5 runs ($\chi 2 = 2.01$ for the best fit) which indicates, that the former epitope location is preferable. The SAXS data and the rigid body model of the CD44 HABD – scFv MEM-85 complex can be accessed in the Small Angle Scattering Biological Databank: http://www.sasbdb.org/data/SASDAG7/.

Finally, modeling of the entire structure with the flexible C-terminus of CD44 HABD (residues 164-178) was performed using CORAL[31, 111] with the same contact conditions as for the SASREF [31] rigid body modeling. The program was run five times to check the consistency of the results. The resulting models only differed in the position of the flexible tail part. The rigid part of CD44 HABD in the CORAL models was only slightly rotated compared to the models created without the tail part, but the mutual orientation of the molecules remained the same. This indicates that the position of the flexible tail has no major influence on the overall structure of the complex. Overall, the SAXS measurements, data analysis and modelling performed here allowed us to define the mutual location of the CD44 HABD – scFv MEM-85 complex components.

5.4. E7 HPV Disordered protein

This project is done in collaboration with Dr. R.Pierattelli group from Centro di Ricerca di Risonanze Magnetiche (CERM), Florence, Italy.

E7 protein is one of the two primary oncoproteins of high risk human papillomavirus (HPV) types [112]. The 12 kDa protein consists of two parts of approximately equal size: a folded (CR3 domain) and a highly disordered domain. The scattering curve, pair distance distribution function and Kratky plot are shown in Fig. 26 and the results are presented in Table 8. The Kratky plot shows that the protein partially unfolded, and EOM was used for further analysis.

As no high-resolution information for the structured CR3 domain of HPV 16 E7 oncoprotein was available, its in-silico model was used instead, based on the information on the molecular assembly derived from both the NMR (pdb ID: 2F8B) and the X-ray (pdb ID: 2B9D)

structure resolved for the E7 of HPV 45 and 1 correspondingly. The determined molecular mass of the solute suggests that there are different oligomers present in solution, probably dimers and tetramers. In order to model the mixture of these oligomers two pools of 10000 models each were generated by EOM. Multiple runs of the genetic algorithm were performed and the results were averaged to provide quantitative information about the oligomeric states and flexibility of the protein in solution. These results suggest the presence of flexible dimeric (85 ± 5%) as well as tetrameric (15 ± 5%) E7 in solution.

The presence in solution of different forms of E7 (monomer/dimer/tetramer) has been already proposed based on sedimentation experiments [113]. NMR results were also pointing to a heterogeneous system, suggesting a high aggregation tendency for the C-terminal part of the protein, still in the presence of a highly mobile and flexible N-terminal part. Further analysis of the E7 experimental data including the scattering curves of CR3 domain are in progress and will be used to draw final conclusions.

Table 8. SAXS data collection and scattering parameters for HPV 16 E7 protein

Structural parameters	
$I(0)$ (relative) (from P(r))	482
R_g (Å) (from $P(r)$)	35 ± 2
$I(0)$ (cm−1) (from Guinier)	477
R_g (Å) (from Guinier)	34 ± 2
D_{max} (Å)	120 ± 5
Porod volume estimate (Å^3)	57111
Excluded volume estimate (Å^3)	74825
Molecular mass determination	
Molecular mass *MM* (Da) from Porod volume ($V_p * 0.6$)	34300 ± 4000
Molecular mass *MM* (Da) from excluded volume (V_{ex} / 2)	37400 ± 4000
Calculated monomeric *MM* from sequence (Da)	12087

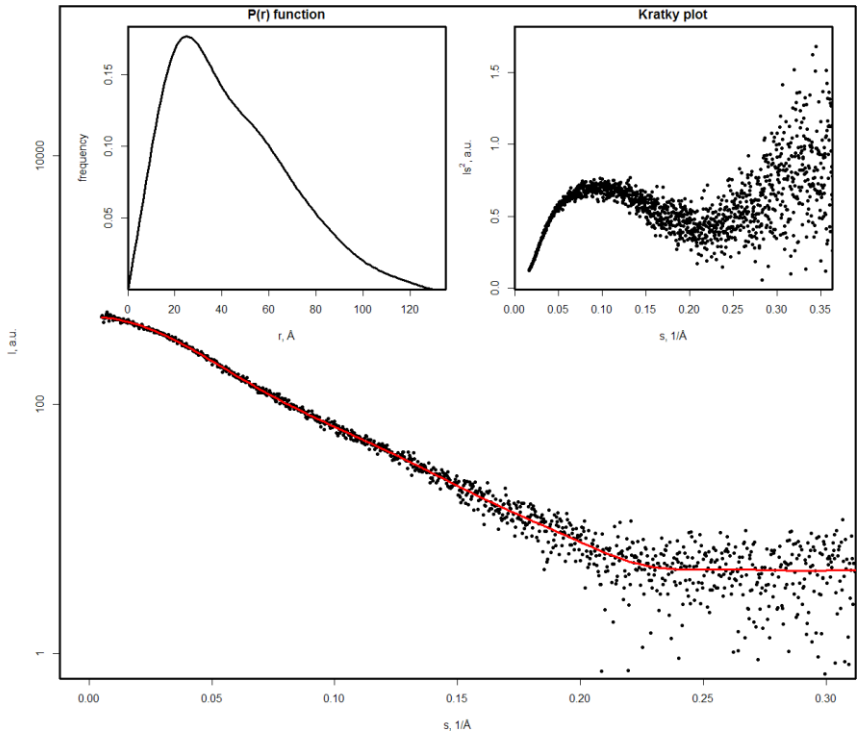

Fig. 26. Experimental SAXS data of the E7 oncoprotein and theoretical scattering from EOM determined ensemble (red). The logarithm of the scattering intensity is plotted against the momentum transfer, using PRIMUS. The figure also shows the derived pair-distance distribution function p(r) in top-left corner and Kratky plot in top-right.

5.5. RTX domain of CyaA protein

This project is presented in the manuscript "Calcium-dependent folding directs protein translocation through Type I secretion system by an intramolecular ratchet mechanism" by Ladislav Bumba, Jiri Masin, Pavel Macek, Tomas Wald, Lucia Motlova, Ilona Bibova, Nela Klimova, Lucie Bednarova, Michael Kachala, Dmitri I. Svergun, Cyril Barinka, Peter Sebo.

The type I secretion system (T1SS) is one of the simplest secretion apparatus of Gram-negative bacteria mediating secretion of proteins directly from the cytoplasm to the extracellular space and most of the T1SS substrates belong to a Repeats-in-toxin (RTX) family of exoproteins with various biological activities [114, 115]. In this study a *Bordetella pertussis* adenylate cyclase toxin (CyaA) was used as a model system (Fig. 27A). SAXS was used in this project characterize one RTX block $CyaA_{1530-1680}$ as well as the entire RTX domain $CyaA_{1009-1706}$ (consisting of five blocks) of CyaA, obtained structural parameters are shown in the Table 9 and $p(r)$ function for entire domain is shown in Fig. 27C. The low resolution *ab initio* models of entire

RTX domain were calculated using the program DAMMIF [28]. The algorithm employs simulated annealing to determine the shape of the macromolecule represented as densely packed beads within a sphere with a diameter (D_{max}) by minimizing the discrepancy χ^2 between experimental scattering curve and the one corresponding to the model. For each construct 20 initial models were calculated and then clustered and averaged with program DAMCLUST [8]. *Ab initio* model of the entire RTX domain revealed an elongated molecule divided into five distinct domains (Fig. 27E). Superimposition of the $CyaA_{1530-1680}$ crystal structure onto the low resolution model showed that the size and shape of $CyaA_{1530-1680}$ correspond well to the structure at the tip of the entire RTX domain suggesting that each of the five domains may correspond to the individual RTX repeat block adopting a β-roll structure. These results allowed authors to propose a model for the translocation of even large RTX proteins by the T1SS apparatus.

Table 9. Overall results of the SAXS experiment for the RTX domain of CyaA

RTX domain construct	Block I-V (1009-1706) 5 repeat blocks	Block V (CyaA1530-1680) 1 repeat block
Structural parameters		
$I(0)$ (A.U.) [from $P(r)$]	17070	2430
R_g (nm) [from $P(r)$]	5.49	1.8
$I(0)$ (A.U.) [from Guinier]	17551	2394
R_g (nm) [from Guinier]	5.55	1.7
D_{max} (nm)	17.15	5.9
Porod volume estimate (nm^3)	275	24
Molecular mass determination		
Partial specific volume ($cm^3\ g^{-1}$)	0.724	0.724
Contrast ($\Delta\rho \times 10^{10}\ cm^{-2}$)	3.047	3.047
Molecular mass (kDa) [from $I(0)$]	91	13
Molecular mass (kDa) [from Porod volume]	162	14
Calculated monomeric molecular mass from sequence	70	16

5.6. Conclusion

The collaborative projects presented in this chapter demonstrate SAXS applications to structural characterization of a wide range of biological macromolecules. The joint use of data obtained with various biological and biophysical methods allows a comprehensive structural characterization of the proteins. Application of *ab initio* methods provides the overall shape of the particles and, in case of multiphase modelling, of protein subunits. Rigid body modelling yields information about mutual location of the components in a complex and may further help in confirming the binding regions. Finally, application of EOM makes it possible to qualitatively characterize flexible proteins. All used methods are part of the SAXS data analysis package ATSAS [6-8], a powerful tool for the structural characterization of macromolecular solutions of proteins and functional complexes.

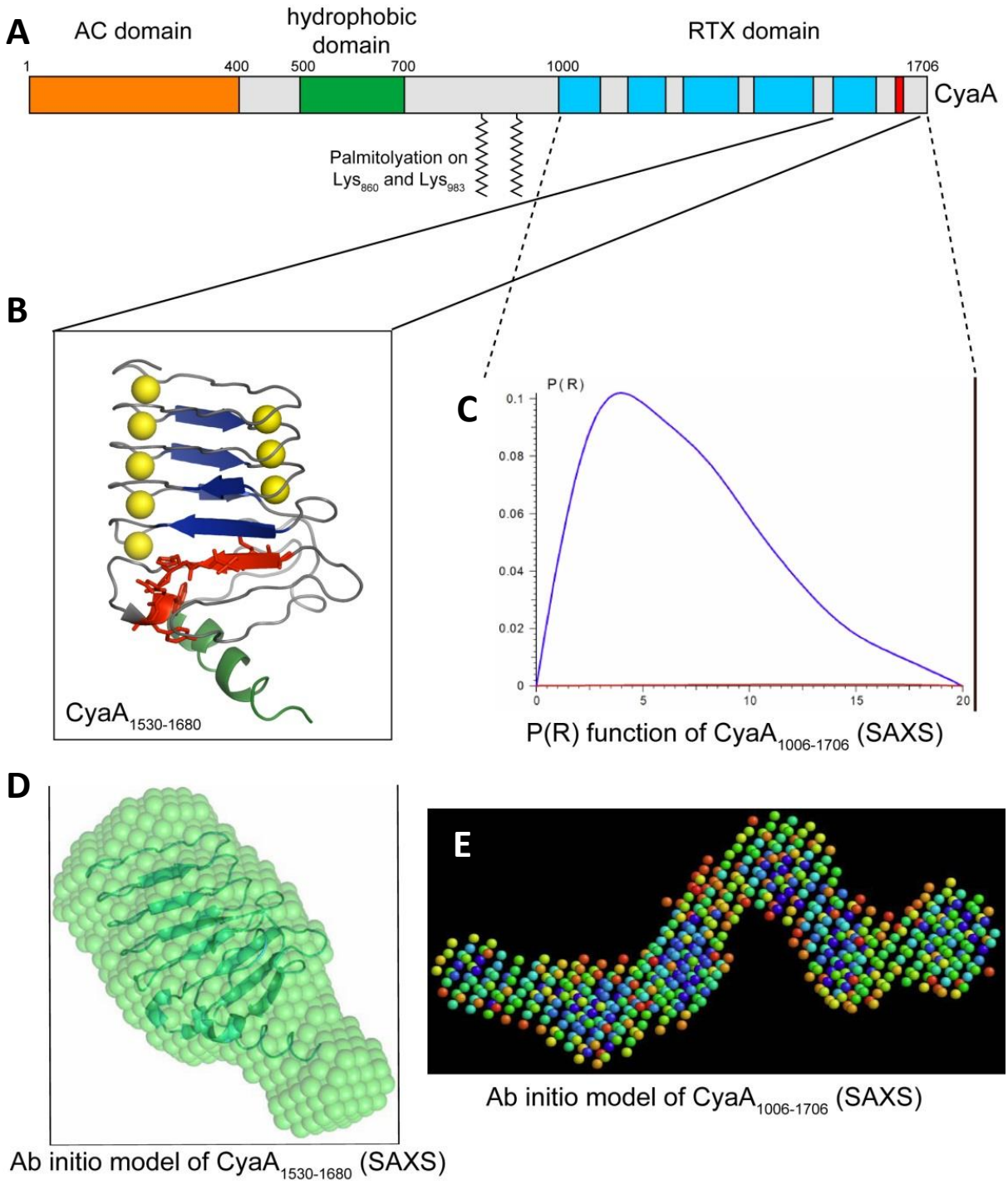

Fig. 27. Results of SAXS measurements and modelling of RTX domain of CyaA. A. Scheme of CyaA protein. B. Crystal structure of one RTX repeat CyaA$_{1530-1680}$. C. Pair-distance distribution function of the entire RTX domain. D. Superimposition of high-resolution and ab initio models of CyaA$_{1530-1680}$. E. Ab initio model of the entire RTX domain.

Conclusions

In recent years due to advances in instrumentation and data analysis approaches small angle scattering became one of the most widely applied techniques for structural characterization of biological macromolecules. The increased amounts of SAS data, diversity and complexity of the applications prompted a need in further development of advanced data analysis and archiving methods. In the PhD project presented in this work some of the acute issues in the field were considered and the solutions were proposed and developed.

The extension of sasCIF format presented in Chapter 2 is an important step towards standardization of the representation and exchange of SAS data and the SAS-based models. The sasCIFtools developed for the processing of sasCIF files facilitate the conversion of the conventional SAS data files and models to the updated sasCIF and *vice versa*. Therefore all kinds of information used in SAS data analysis (scattering patterns, distance distribution functions, models and fits of their calculated scattering to the experimental data) can be included in a single sasCIF file and easily exchanged. The integration of the sasCIFtools into the SASBDB database [9] opens the possibility of online export and import of the entire database entries as one sasCIF file. Following the wwPDB small-angle scattering task force recommendations [11], these tools facilitate data exchange between SAS federated databases. These measures together with the introduction of SAS databases make the data organization and management more accessible for users and promotes SAS applications in the structural biology community.

A practically important problem of characterization of solutions with interparticle interactions is addressed by the development of a Monte-Carlo based algorithm (Chapter 3) for simultaneous determination of form and structure factors. The proposed algorithm was shown to reconstruct structure factor contributions for interacting systems and upon its planned introduction to the ATSAS package the program will become publically available.

Characterization of structural properties of disordered proteins and other flexible structures is a topic of great interest in structural biology today and SAXS is one of the most powerful and widely applied techniques for the analysis of such objects. The ensemble approach is an optimal way to quantitatively characterize their properties in solution and the first implementation of an ensemble approach for SAXS was the Ensemble Optimization Method (EOM), recently updated to a new version, EOM 2.0. A series of tests and case studies conducted within

the scope of this project (Chapter 4) have shown that EOM 2.0 is able to correctly represent the properties of the unfolded proteins, resolve subpopulations in mixtures and is robust to the noise in the experimental data.

Finally, applications of SAXS for characterization of diverse proteins as part of collaborative projects with EMBL beamline users are presented in Chapter 5. The employed data analysis and modeling methods include *ab initio*, rigid body modelling and EOM. The *ab initio* approach was applied to determine the overall shape of the globular proteins in each of the projects and multiphase *ab initio* models helped to define the mutual location of the I27-PimA fusion protein subunits. Rigid body modelling revealed the relative positions of the components of CD44 MEM-85 complex and validated the location of the epitiope and paratope suggested by other methods. The application of EOM to the disordered protein E7 allowed us to quantitatively characterize flexibility and oligomeric state of the protein.

References

1. Svergun, D.I., et al., *Small Angle X-Ray and Neutron Scattering from Solutions of Biological Macromolecules*. 2013: OUP Oxford.
2. Guinier, A., *La diffraction des rayons X aux tres petits angles; application a l'etude de phenomenes ultramicroscopiques.* Ann. Phys. (Paris), 1939. **12**: p. 161-237.
3. Ritland, H., P. Kaesberg, and W. Beeman, *An X-Ray Investigation of the Shapes and Hydrations of Several Protein Molecules in Solution.* The Journal of Chemical Physics, 1950. **18**(9): p. 1237-1242.
4. Feigin, L.A. and D.I. Svergun, *Structure analysis by small-angle x-ray and neutron scattering*. 1987, New York: Plenum Press. xiii, 335.
5. Jeffries, C.M. and D.I. Svergun, *High-throughput studies of protein shapes and interactions by synchrotron small-angle x-ray scattering.* Methods Mol Biol, 2015. **1261**: p. 277-301.
6. Konarev, P.V., et al., *ATSAS 2.1, a program package for small-angle scattering data analysis.* J. Appl. Crystallogr., 2006. **39**: p. 277-286.
7. Petoukhov, M.V., et al., *ATSAS 2.1 - towards automated and web-supported small-angle scattering data analysis.* J. Appl. Cryst., 2007. **40**(s1): p. s223-s228.
8. Petoukhov, M.V., et al., *New developments in the ATSAS program package for small-angle scattering data analysis.* Journal of Applied Crystallography, 2012. **45**(2): p. 342-350.
9. Valentini, E., et al., *SASBDB, a repository for biological small-angle scattering data.* Nucleic Acids Res, 2015. **43**(Database issue): p. D357-63.
10. Hura, G.L., et al., *Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS).* Nat Methods, 2009. **6**(8): p. 606-12.
11. Trewhella, J., et al., *Report of the wwPDB Small-Angle Scattering Task Force: Data Requirements for Biomolecular Modeling and the PDB.* Structure, 2013. **21**(6): p. 875-881.
12. Malfois, M. and D.I. Svergun, *SasCIF - an extension of core Crystallographic Information File for small angle scattering.* J. Appl. Crystallogr., 2000. **34**: p. 812-816.
13. Svergun, D.I., *Determination of the regularization parameter in indirect-transform methods using perceptual criteria.* J. Appl. Crystallogr., 1992. **25**: p. 495-503.
14. Bernado, P. and D.I. Svergun, *Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering.* Mol Biosyst, 2012. **8**(1): p. 151-67.
15. Bernado, P., et al., *Structural characterization of flexible proteins using small-angle X-ray scattering.* J Am Chem Soc, 2007. **129**(17): p. 5656-64.
16. Tria, G., et al., *Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering.* IUCrJ, 2015. **2**(2): p. 207-217.
17. Kratky, O., *X-ray Small Angle Scattering with Substances of Biological Interest in Diluted Solutions*. 1963: Pergamon Press.
18. Svergun, D.I. and M.H.J. Koch, *Small angle scattering studies of biological macromolecules in solution.* Rep. Progr. Phys., 2003. **66**: p. 1735-1782.
19. Svergun, D.I., *Small-angle X-ray and neutron scattering as a tool for structural systems biology.* Biol Chem, 2010. **391**(7): p. 737-43.
20. Glatter, O., *A new method for the evaluation of small-angle scattering data.* J. Appl. Cryst., 1977. **10**: p. 415-421.

21. Semenyuk, A.V. and D.I. Svergun, *GNOM - a program package for small-angle scattering data processing.* J. Appl. Crystallogr., 1991. **24**: p. 537-540.
22. Porod, G., *General theory*, in *Small-angle X-ray scattering*, O. Glatter and O. Kratky, Editors. 1982, Academic Press: London. p. 17-51.
23. Glatter, O. and O. Kratky, *Small Angle X-ray Scattering*. 1982, London: Academic Press. 515.
24. Stuhrmann, H.B., *Interpretation of small-angle scattering functions of dilute solutions and gases. A representation of the structures related to a one-particle-scattering function.* Acta Cryst., 1970. **A26**: p. 297-306.
25. Blanchet, C.E. and D.I. Svergun, *Small-angle X-ray scattering on biological macromolecules and nanocomposites in solution.* Annu Rev Phys Chem, 2013. **64**: p. 37-54.
26. Chacon, P., et al., *Low-resolution structures of proteins in solution retrieved from X-ray scattering with a genetic algorithm.* Biophys J, 1998. **74**(6): p. 2760-75.
27. Svergun, D.I., *Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing.* Biophys J, 1999. **76**(6): p. 2879-86.
28. Franke, D. and D.I. Svergun, *DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering.* J. Appl. Cryst., 2009. **42**: p. 342-346.
29. Svergun, D.I., C. Barberato, and M.H.J. Koch, *CRYSOL - a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates.* J. Appl. Crystallogr., 1995. **28**: p. 768-773.
30. Svergun, D.I., et al., *Protein hydration in solution: experimental observation by x-ray and neutron scattering.* Proc Natl Acad Sci U S A, 1998. **95**(5): p. 2267-72.
31. Petoukhov, M.V. and D.I. Svergun, *Global rigid body modeling of macromolecular complexes against small-angle scattering data.* Biophys J, 2005. **89**(2): p. 1237-1250.
32. Graewert, M.A. and D.I. Svergun, *Impact and progress in small and wide angle X-ray scattering (SAXS and WAXS).* Curr Opin Struct Biol, 2013. **23**(5): p. 748-54.
33. Hura, G.L., et al., *Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS).* Nature Methods, 2009. **6**(8): p. 606-U83.
34. Hall, S.R., F.H. Allen, and I.D. Brown, *The crystallographic information file (CIF): a new standard archive file for crystallography.* Acta Crystallographica Section A: Foundations of Crystallography, 1991. **47**(6): p. 655-685.
35. Westbrook, J.D. and P.E. Bourne, *STAR/mmCIF: an ontology for macromolecular structure.* Bioinformatics, 2000. **16**(2): p. 159-168.
36. Hall, S.R. and A.P.F. Cook, *Star Dictionary Definition Language - Initial Specification.* Journal of Chemical Information and Computer Sciences, 1995. **35**(5): p. 819-825.
37. Fitzgerald, P., et al. *The macromolecular CIF dictionary*. in *American Crystallographic Association Annual Meeting, Albuquerque, NM USA*. 1993.
38. Hall, S.R., *The Star File - a New Format for Electronic Data Transfer and Archiving.* Journal of Chemical Information and Computer Sciences, 1991. **31**(2): p. 326-333.
39. Hall, S.R. and N. Spadaccini, *The Star File - Detailed Specifications.* Journal of Chemical Information and Computer Sciences, 1994. **34**(3): p. 505-508.
40. Spadaccini, N. and S.R. Hall, *Extensions to the STAR File Syntax.* Journal of Chemical Information and Modeling, 2012. **52**(8): p. 1901-1906.
41. Westbrook, J. and S. Hall, *A dictionary description language for macromolecular structure.* Rutgers University, New Brunswick, NJ, Report NDB-110, 1995.

42. Homan, E., et al., *The SAXS/WAXS software system of the DUBBLE CRG beamline at the ESRF*. J. Appl. Cryst., 2001. **34**: p. 519-522.
43. Jacques, D.A., et al., *Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution.* Acta Crystallographica Section D, 2012. **68**(6): p. 620-626.
44. Franke, D., C.M. Jeffries, and D.I. Svergun, *Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra.* Nat Meth, 2015. **advance online publication**.
45. Ginkel, G.v. *PDBeCIF*. [cited 2015 01.02.2015]; Available from: http://www.ebi.ac.uk/~glen/PDBeCIF/.
46. Franke, D. *saxsview Project*. [cited 2015 01.02.2015]; Available from: http://saxsview.sourceforge.net/.
47. Dustman, A., *MySQLdb User's Guide*.
48. Franke, D., A.G. Kikhney, and D.I. Svergun, *Automated acquisition and analysis of small angle X-ray scattering data.* Nuclear Instruments and Methods in Physics Research A, 2012. **689**: p. 52–59.
49. Pedersen, J.S., *Modelling of small-angle scattering data from colloids and polymer systems*. 2002: North Holland, Elsevier: Amsterdam.
50. Brunner-Popela, J. and O. Glatter, *Small-angle scattering of interacting particles. I. Basic principles of a global evaluation technique.* J. Appl. Cryst., 1997. **30**: p. 431-442.
51. Bergmann, A., G. Fritz, and O. Glatter, *Solving the generalized indirect Fourier transformation (GIFT) by Boltzmann simplex simulated annealing (BSSA).* Journal of applied crystallography, 2000. **33**(5): p. 1212-1216.
52. Weyerich, B., J. Brunner-Popela, and O. Glatter, *Small-angle scattering of interacting particles. II. Generalized indirect Fourier transformation under consideration of the effective structure factor for polydisperse systems.* J. Appl. Cryst., 1999. **32**: p. 197-209.
53. Svergun, D.I., A.V. Semenyuk, and L.A. Feigin, *Small angle scattering data treatment by the regularization method.* Acta Crystallogr. (A), 1988. **44**: p. 244-250.
54. Tikhonov, A.N. *On the stability of inverse problems*. in *Dokl. Akad. Nauk SSSR*. 1943.
55. Fukasawa, T. and T. Sato, *Versatile application of indirect Fourier transformation to structure factor analysis: from X-ray diffraction of molecular liquids to small angle scattering of protein solutions.* Phys Chem Chem Phys, 2011. **13**(8): p. 3187-96.
56. Ornstein, L.S. and F. Zernike. *Accidental deviations of density and opalescence at the critical point of a single substance*. in *Proc. Akad. Sci.(Amsterdam)*. 1914.
57. Percus, J.K. and G.J. Yevick, *Analysis of classical statistical mechanics by means of collective coordinates.* Physical Review, 1958. **110**(1): p. 1.
58. Perram, J., *Hard sphere correlation functions in the Percus-Yevick approximation.* Molecular Physics, 1975. **30**(5): p. 1505-1509.
59. Pedersen, J.S., *Determination of size distribution from small-angle scattering data for systems with effective hard-sphere interactions.* Journal of applied crystallography, 1994. **27**(4): p. 595-608.
60. Hansen, S., *Simultaneous estimation of the form factor and structure factor for globular particles in small-angle scattering.* Journal of Applied Crystallography, 2008. **41**(2): p. 436-445.
61. Tikhonov, A. and V.Y. Arsenin, *Solutions of ill-posed problems.* WH Winston, Washington, DC, 1977. **330**.

62. Eaton, J.W., D. Bateman, and S. Hauberg, *GNU Octave version 3.0. 1 manual: a high-level interactive language for numerical computations*. 2007: SoHo Books.
63. Kinning, D.J. and E.L. Thomas, *Hard-sphere interactions between spherical domains in diblock copolymers.* Macromolecules, 1984. **17**(9): p. 1712-1718.
64. Uversky, V.N. and A.K. Dunker, *Understanding protein non-folding.* Biochim Biophys Acta, 2010. **1804**(6): p. 1231-64.
65. Dunker, A.K. and Z. Obradovic, *The protein trinity--linking function and disorder*. Nat Biotechnol, 2001. **19**(9): p. 805-6.
66. Uversky, V.N., C.J. Oldfield, and A.K. Dunker, *Intrinsically disordered proteins in human diseases: introducing the D2 concept.* Annu. Rev. Biophys., 2008. **37**: p. 215-246.
67. Rezaei-Ghaleh, N., M. Blackledge, and M. Zweckstetter, *Intrinsically disordered proteins: from sequence and conformational properties toward drug discovery.* Chembiochem, 2012. **13**(7): p. 930-50.
68. Dunker, A.K., et al., *Intrinsic protein disorder in complete genomes.* Genome Inform Ser Workshop Genome Inform, 2000. **11**: p. 161-71.
69. Uversky, V.N., J.R. Gillespie, and A.L. Fink, *Why are "natively unfolded" proteins unstructured under physiologic conditions?* Proteins, 2000. **41**(3): p. 415-27.
70. Schweers, O., et al., *Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure.* J Biol Chem, 1994. **269**(39): p. 24290-7.
71. Dunker, A.K., et al., *Intrinsically disordered protein.* J Mol Graph Model, 2001. **19**(1): p. 26-59.
72. Dunker, A.K., et al., *Intrinsic disorder and protein function.* Biochemistry, 2002. **41**(21): p. 6573-82.
73. Xue, B., et al., *PONDR-FIT: a meta-predictor of intrinsically disordered amino acids.* Biochim Biophys Acta, 2010. **1804**(4): p. 996-1010.
74. Dosztanyi, Z., et al., *IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.* Bioinformatics, 2005. **21**(16): p. 3433-4.
75. Jensen, M.R., R.W. Ruigrok, and M. Blackledge, *Describing intrinsically disordered proteins at atomic resolution by NMR.* Current opinion in structural biology, 2013. **23**(3): p. 426-435.
76. Gadella, T.W.J., *FRET and FLIM Techniques*. 2011: Elsevier Science.
77. Schuler, B., et al., *Application of confocal single-molecule FRET to intrinsically disordered proteins,* in *Intrinsically Disordered Protein Analysis*. 2012, Springer. p. 21-45.
78. Konarev, P.V., et al., *PRIMUS - a Windows-PC based system for small-angle scattering data analysis.* J. Appl. Crystallogr., 2003. **36**: p. 1277-1282.
79. Bernado, P. and M. Blackledge, *A self-consistent description of the conformational behavior of chemically denatured proteins from NMR and small angle scattering.* Biophys J, 2009. **97**(10): p. 2839-45.
80. Flory, P.J., *Principles of Polymer Chemistry*. 1953: Cornell University Press.
81. Le Guillou, J.C. and J. Zinn-Justin, *Critical Exponents for the n-Vector Model in Three Dimensions from Field Theory.* Physical Review Letters, 1977. **39**(2): p. 95-98.
82. Kohn, J.E., et al., *Random-coil behavior and the dimensions of chemically unfolded proteins.* Proc Natl Acad Sci U S A, 2004. **101**(34): p. 12491-6.

83. Stumpe, M.C. and H. Grubmüller, *Interaction of urea with amino acids: implications for urea-induced protein denaturation.* Journal of the American Chemical Society, 2007. **129**(51): p. 16126-16131.
84. Meier, S., S. Grzesiek, and M. Blackledge, *Mapping the conformational landscape of urea-denatured ubiquitin using residual dipolar couplings.* J Am Chem Soc, 2007. **129**(31): p. 9799-807.
85. Bernado, P., et al., *A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering.* Proc Natl Acad Sci U S A, 2005. **102**(47): p. 17002-7.
86. von Ossowski, I., et al., *Protein disorder: conformational distribution of the flexible linker in a chimeric double cellulase.* Biophys J, 2005. **88**(4): p. 2823-32.
87. Tompa, P., *On the supertertiary structure of proteins.* Nat Chem Biol, 2012. **8**(7): p. 597-600.
88. Pelikan, M., G.L. Hura, and M. Hammel, *Structure and flexibility within proteins as identified through small angle X-ray scattering.* Gen Physiol Biophys, 2009. **28**(2): p. 174-89.
89. Yang, S., et al., *Multidomain assembled states of Hck tyrosine kinase in solution.* Proc Natl Acad Sci U S A, 2010. **107**(36): p. 15757-62.
90. Rozycki, B., Y.C. Kim, and G. Hummer, *SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions.* Structure, 2011. **19**(1): p. 109-16.
91. Krzeminski, M., et al., *Characterization of disordered proteins with ENSEMBLE.* Bioinformatics, 2013. **29**(3): p. 398-9.
92. Mylonas, E., et al., *Domain conformation of tau protein studied by solution small-angle X-ray scattering.* Biochemistry, 2008. **47**(39): p. 10345-10353.
93. Bernado, P., et al., *Structure and Dynamics of Ribosomal Protein L12: An Ensemble Model Based on SAXS and NMR Relaxation.* Biophys J, 2010. **98**(10): p. 2374-82.
94. Shkumatov, A.V., et al., *Structural memory of natively unfolded tau protein detected by small-angle X-ray scattering.* Proteins, 2011. **79**(7): p. 2122-31.
95. Gatzeva-Topalova, P.Z., et al., *Structure and flexibility of the complete periplasmic domain of BamA: the protein insertion machine of the outer membrane.* Structure, 2010. **18**(11): p. 1492-1501.
96. Ellis, J., et al., *Domain motion in cytochrome P450 reductase conformational equilibria revealed by NMR and small-angle X-ray scattering.* Journal of Biological Chemistry, 2009. **284**(52): p. 36628-36637.
97. Team, R., *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria*. 2008.
98. Fitzkee, N.C. and G.D. Rose, *Reassessing random-coil statistics in unfolded proteins.* Proc Natl Acad Sci U S A, 2004. **101**(34): p. 12497-502.
99. Tanford, C., K. Kawahara, and S. Lapanje, *Proteins in 6-M guanidine hydrochloride. Demonstration of random coil behavior.* J Biol Chem, 1966. **241**(8): p. 1921-3.
100. Hong, L. and J. Lei, *Scaling law for the radius of gyration of proteins and its dependence on hydrophobicity.* Journal of Polymer Science Part B: Polymer Physics, 2009. **47**(2): p. 207-214.
101. Blanchet, C., et al., *Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY).* Journal of Applied Crystallography, 2015. **48**(2): p. 0-0.

102. Kordulakova, J., et al., *Definition of the First Mannosylation Step in Phosphatidylinositol Mannoside Synthesis PimA IS ESSENTIAL FOR GROWTH OF MYCOBACTERIA.* Journal of Biological Chemistry, 2002. **277**(35): p. 31335-31344.
103. Guerin, M.E., et al., *Molecular basis of phosphatidyl-myo-inositol mannoside biosynthesis and regulation in mycobacteria.* Journal of Biological Chemistry, 2010. **285**(44): p. 33577-33583.
104. Blanchet, C.E., et al., *Instrumental setup for high-throughput small- and wide-angle solution scattering at the X33 beamline of EMBL Hamburg.* Journal of Applied Crystallography, 2012. **45**: p. 489-495.
105. Roessle, M.W., et al., *Upgrade of the small-angle X-ray scattering beamline X33 at the European Molecular Biology Laboratory, Hamburg.* J. Appl. Cryst., 2007. **40**: p. s190-s194.
106. Volkov, V.V. and D.I. Svergun, *Uniqueness of ab initio shape determination in small angle scattering.* J. Appl. Crystallogr., 2003. **36**: p. 860-864.
107. Takeda, M., et al., *Ligand-induced structural changes of the CD44 hyaluronan-binding domain revealed by NMR.* J Biol Chem, 2006. **281**(52): p. 40089-95.
108. Sircar, A., E.T. Kim, and J.J. Gray, *RosettaAntibody: antibody variable region homology modeling server.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W474-9.
109. Kozin, M.B. and D.I. Svergun, *Automated matching of high- and low-resolution structural models.* Journal of Applied Crystallography, 2001. **34**(1): p. 33-41.
110. Bajorath, J., et al., *Identification of CD44 residues important for hyaluronan binding and delineation of the binding site.* J Biol Chem, 1998. **273**(1): p. 338-43.
111. Petoukhov, M.V., et al., *New developments in the program package for small-angle scattering data analysis.* J Appl Crystallogr, 2012. **45**(Pt 2): p. 342-350.
112. Boyer, S.N., D.E. Wazer, and V. Band, *E7 protein of human papilloma virus-16 induces degradation of retinoblastoma protein through the ubiquitin-proteasome pathway.* Cancer research, 1996. **56**(20): p. 4620-4624.
113. Clements, A., et al., *Oligomerization properties of the viral oncoproteins adenovirus E1A and human papillomavirus E7 and their complexes with the retinoblastoma protein.* Biochemistry, 2000. **39**(51): p. 16033-45.
114. Thomas, S., I.B. Holland, and L. Schmitt, *The Type 1 secretion pathway - the hemolysin system and beyond.* Biochim Biophys Acta, 2014. **1843**(8): p. 1629-41.
115. Linhartova, I., et al., *RTX proteins: a highly diverse family secreted by a common mechanism.* FEMS Microbiol Rev, 2010. **34**(6): p. 1076-112.

Appendix. Eidesstattliche Versicherung

Declaration on oath

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

I hereby declare, on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.

Hamburg, den 10.04.2015