Software Development for High-throughput Biological Small Angle X-ray Scattering

Dissertation
zur Erlangung des Doktorgrades der Naturwissenschaften
am Department für Chemie
der Universität Hamburg

vorgelegt von
Alexey Kikhney
aus Wladiwostok

Hamburg
2010

Gutachter:

Prof. Dr. Dr. Christian Betzel

Prof. Dr. Ulrich Hahn

Tag der Disputation: 23. Juli 2010

Abstract

Small angle X-ray scattering (SAXS) is a fundamental tool in the study of biological macromolecules providing low resolution information on the shape, conformation, assembly state and folding/unfolding of macromolecules in solution. The possibility to rapidly study the structure of proteins, nucleic acids and their complexes in near physiological environments makes SAXS very important for structural biology, despite the limited resolution range. Recent advances in robotic microliter fluid handling make it possible to perform automated high throughput experiments including fast screening of solution conditions, structural responses to ligand binding, changes in temperature and chemical modifications. However, in order to make automated experiments a routine tool suitable for large scale structural studies, several significant bottlenecks must be overcome. Among the most critical ones are the issues of reliability of the experiment, automation of data analysis, control of the data quality and convenient storage of the obtained results.

The present study addresses acute problems in the automation of modern SAXS experiments and in the methods for rapid data processing, analysis and storage. A new visualization hardware setup was designed for the automated control of the sample cell filling, coupled with an image recognition program to detect and correct for improper filling. Software was developed and implemented to assess the quality of the scattering data and automatically compute a classical SAXS parameter, the radius of gyration. A pipeline that integrates multiple software modules, both existing and newly developed, was implemented to make data processing possible without user intervention. The pipeline includes: intelligent background subtraction, analysis of concentration effects, calculation of overall parameters and characteristic functions, low resolution *ab initio* shape reconstruction and the use of a SAXS patterns database. A cross-validation of the intermediate results ensures a robust reproducible outcome which is stored in a user-friendly web-compatible XML format. The data analysis pipeline allows introduction of further decision making blocks and extensive use of *a priori* information, paving the way for development of an intelligent expert system for SAXS-based model building.

The developed hardware and software was implemented at the beamline X33 of EMBL Hamburg and, in combination with the two robotic sample changers, improved dramatically the reliability, throughput and user friendliness of the beamline, ensuring its fully automated operation. The automated setup was successfully utilized by over 200 user groups in 2008-2009. The applications of these methods are illustrated in this work in several collaborative user projects on different macromolecular systems.

Zusammenfassung

Röntgenkleinwinkelstreuung (small angle X-ray scattering, SAXS) ist ein fundamentelles Werkzeug zur Untersuchung biologischer Makromoleküle in Lösung. Diese Methode ermöglicht die Untersuchung der Form, Konformation, des Aufbaus und der Faltung/Entfaltung von Makromolekülen bei niedriger Auflösung. Trotz der begrenzten Auflösung sind SAXS Studien im Bereich der Strukturbiologie essentiell, da sie strukturelle Informationen über Proteine, Nukleinsäuren und deren Komplexe unter annähernd physiologischen Bedingungen liefern.

Neueste Fortschritte im Gebiet der automatisierten Handhabung von Flüssigkeiten im Milliliter Bereich ermöglichen heute die Durchführung von Hochdurchsatzexperimenten wie die schnelle Untersuchung von Lösungmittelsbedingungen, die strukturellen Auswirkungen der Bindung von Liganden, Temperaturveränderungen, chemische Modifikationen, etc..

Allerdings ist die routinemäßige Anwendung automatisierter Experimente im großen Maßstab in der Strukturbiologie bis heute nur eingeschränkt möglich. Am kritischsten in diesem Zusammenhang sind Grenzen in der experimentellen Verlässlichkeit, bei der automatischen Datenanalyse sowie der kontrollierten Datenqualität und bei der Datenspeicherung.

In der hier vorgelegten Arbeit werden Lösungen für diese Schwierigkeiten bei der Automatisierung moderner SAXS Experimente erarbeitet und Methoden für eine schnelle Datenverarbeitung, -analyse und -speicherung entwickelt. Für die automatische Kontrolle der Probenkammerbefüllung wurde ein neues Visualisierungssystem entworfen, das mit einem Bilderkennungsprogramm zur Erkennung und Korrektur fehlerhafter Kammerbefüllung gekoppelt ist. Zudem wurde ein Computerprogramm zur automatischen Berechnung klassischer SAXS Parameter wie z.B. des Trägheitsradius (radius of gyration, R_g) entwickelt und etabliert. Dieses Programm wurde so gestaltet, dass zusätzlich automatisch eine Einschätzung der Datenqualität erfolgt.

Um eine Datenverarbeitung ohne manuelle Eingriffe der Anwender zu ermöglichen wurde eine Programm-Pipeline eingeführt, die mehrere neu entwickelte und bereits

bestehende Programmmodule in sich vereinigt. Diese Pipeline beinhaltet eine kontrollierte Subtraktion von Hintergrundstreuung, eine Analyse von Konzentrationseffekten sowie die Berechnung von allgemeinen Parametern und charakteristischen Funktionen. Zudem ermöglicht diese Pipeline die Rekonstruktion der *ab initio* Form des zu untersuchenden Moleküls bei niedriger Auflösung und den Gebrauch einer SAXS Musterdatenbank. Eine Vergleichsprüfung der Zwischenergebnisse stellt robuste, reproduzierbare Resultate sicher, die in einem anwenderfreundlichen webkompatiblen XML Format gestaltet sind. Der Abschnitt der Pipeline, der für die Datenanalyse zuständig ist, erlaubt die Einführung weiterer Entscheidungsblöcke und die intensive Nutzung von *a priori* Informationen. Diese Ergebnisse schaffen die Grundlage für die weitere Entwicklung eines intelligenten Expertensystems für das SAXS-basierte Modellieren von Strukturen.

Die entwickelten Systeme und Programme wurden in die Beamline X33 des EMBL Hamburg integriert. Dies erhöhte in Kombination mit zwei automatischen Probenaustauschsystemen drastisch die Zuverlässigkeit, den Probenumsatz und die Anwenderfreundlichkeit der Beamline und ermöglichte erstmals eine vollautomatische Bedienung. Das automatisierte System wurde bereits von über 200 Nutzergruppen in einem Zeitraum von 2008-2009 erfolgreich genutzt. In dieser Arbeit wird die Anwendung der automatisierten Methoden beispielhaft an verschiedenen Makromolekülsystemen aus mehreren Kooperationen mit Nutzerprojekten illustriert.

Table of contents

List of figures

List of tables

List of abbreviations

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
AUC	Analytical ultracentrifugation
BMS	Beamline meta server
BSA	Bovine serum albumin
cd	Candela
CSV	Comma-separated values
DLS	Dynamic light scattering
DESY	Deutsches Electronen Synchrotron
D_{max}	Maximum dimension
DTT	Dithiothreitol
EM	Electron microscopy
EPR	Electron paramagnetic resonance
Fig.	Figure
FRET	Fluorescence resonance energy transfer
GUI	Graphical user interface
HTML	Hypertext markup language
HTTP	Hypertext transfer protocol
LED	Light emitting diode
mg	Milligram
ml	Milliliter
nm	Nanometer
NMR	Nuclear magnetic resonance
RDC	Residual dipolar coupling

R_g	Radius of gyration
SANS	Small angle neutron scattering
SAS	Small angle scattering
SAXS	Small angle X-ray scattering
SSV	Space-separated values
TINE	Three-fold integrated networking environment
V_{DAM}	Dummy atom model volume
V_P	Porod volume
XML	Extensible markup language
XSL	Extensible style sheet language

1 Introduction

Small-angle X-ray scattering (SAXS) is a fundamental technique that allows the study of the structure and interactions of proteins, nucleic acids and their complexes in solution. Macromolecular folding, unfolding, aggregation, shape, conformation, and assembly processes can be studied under a variety of conditions, from near-physiological to highly denaturing, although at low resolution of about 1-2 nm, but without the need to crystallize the sample and without the size limitations inherent in NMR and electron microscopy studies. The increasing availability of high-flux third-generation synchrotron radiation sources, recent progress in instrumentation and novel analysis methods allow to retrieve significantly more structural information than previously believed, especially with the use of complementary information from other techniques.

1.1 SAXS theory

Solution small angle X-ray scattering occurs when a monochromatic beam of photons irradiates a sample (a solution in our case) and is scattered. The signal emerges from a difference (contrast) in the average electron density between the solvent and the solute (particles with small inhomogeneities can be considered to have a uniform electron density distribution at low angles/resolutions). To obtain the scattering of a particle the scattering of the pure solvent is subtracted from the one of solution (containing both solvent and solute). For dilute homogeneous and monodisperse solutions without noticeable interactions between the particles the scattering pattern from the sample after solvent subtraction is proportional to the scattering of the particle averaged over all orientations, since the individual particles can be found in all orientations in the solution.

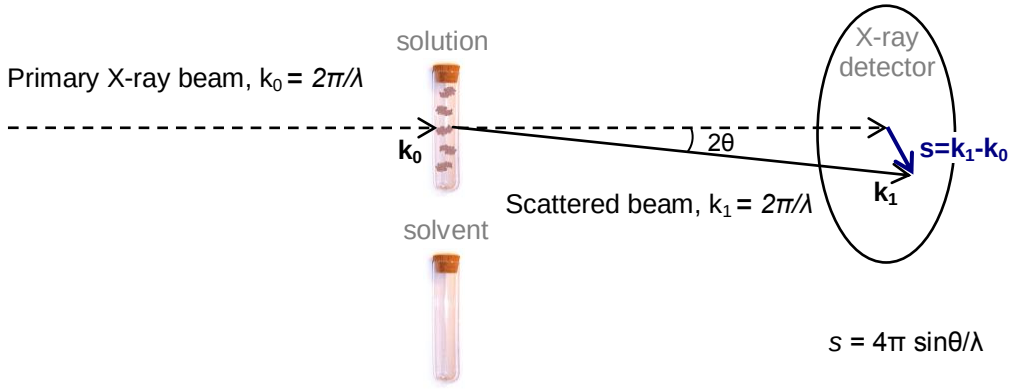


Fig. 1.1 Schematic representation of a SAXS experiment

The scattered intensity I(s) is the Fourier transform of the intramolecular atomic distance distribution function $p(r)$ (i.e. the histogram of the distances within a particle averaged over all orientations). $I(s)$ is a function of the momentum transfer (alternately, modulus of the scattering vector i.e. the directional difference between the incident beam k and the scattered one k_1) $s = 4\pi\sin(\theta)/\lambda$ where 2θ is the scattering angle and λ is the wavelength of the incident beam (Fig. 1.1). Alternately, a letter q is sometimes used instead of s to denote the momentum transfer. The maximum s value in the scattering pattern determines the nominal resolution of the experiment as

$d = 2\pi/s_{max}$. The scattering intensity I(s) is related to the distance distribution function by

$$I(s) = 4\pi \int_0^{D_{\max}} p(r)\frac{\sin(sr)}{sr}dr \qquad \text{Equation 1.1}$$

or, inversely,

$$p(r) = \frac{r^2}{2\pi^2}\int_0^{\infty} \frac{s^2 I(s)\sin(sr)}{sr}ds \qquad \text{Equation 1.2}$$

where D_{max} is the maximum size of the particle. Because of the information loss due to the spherical average, the nominal resolution range of a scattering experiment is not directly related to the resolution of the models obtained. The average leads moreover to the intrinsic ambiguity of the SAXS data, whereby multiple models can fit the data equally well. For some simple objects, e.g. a perfect sphere, such ambiguity does not exist, but real molecules, especially biological macromolecules, have complicated shapes and this ambiguity poses is one of the most important problems when interpreting the SAXS curves.

Though the observed scattering intensity is often measured on a relative scale (depending on the detector collecting the scattered photons) it depends on the size of the particles (larger particles scatter stronger at small angles than smaller ones) as well as the concentration of the particles.

At very low angles the scattering pattern can be described with two parameters which correspond to the molecular mass and the size of the particle. The first parameter, the intensity at zero angle or forward scattering $I(0)$, is proportional to the molecular mass corresponding to in-phase scattering from all electrons in the particle. The second parameter is the radius of gyration (R_g) which corresponds to the averaged squared distance from the center of mass of the particle weighted by the scattering length density. These two parameters are related by the so-called Guinier law (Guinier, 1939) as $I(s) = I(0)\exp[-(sR_g)^2/3)]$ for scattering angles that satisfy the relation

$sR_g < 1.3$ (see section 3.2.1 for details). Forward scattering $I(0)$ cannot be measured directly because it follows the direction of the primary beam, which is masked before reaching the detector by a beamstop. Thus, the Guinier law can be used as an approximation to extrapolate to zero angle. Note that the Guinier law is valid for smaller particles for a longer s-range than for larger ones. This sometimes leads to problems when calculating the $I(0)$ and R_g for larger (or very elongated) particles as the number of experimental points is not sufficient for an accurate estimation. For this reason the relation of $I(s)$ to $p(r)$ is utilized in indirect transform methods (Glatter 1977, Semenyuk & Svergun 1991) and the R_g is calculated from the $p(r)$, essentially using the information from the full scattering curve. While calculating the $p(r)$ directly from $I(s)$ requires integration from zero to infinity and is strongly influenced by the noise in the data, the opposite is much more stable. We do not directly measure $p(r)$ but we can parameterize the $p(r)$ and compute the parameters yielding the computed intensity that agrees with the experimental $I(s)$. This method is especially useful when a Guinier approximation is difficult to obtain.

Another parameter that can be computed from the small angle portion of the data ($s < 0.25$ Å^{-1}) is the excluded volume of the hydrated particle. It is obtained using the Porod equation (Porod 1982)

$$V_P = \frac{2\pi^2 I(0)}{\int_0^\infty [I(s) - K_4] s^2 \, ds} \qquad \text{Equation 1.3}$$

where K_4 is a constant determined to ensure the asymptotical intensity decay proportional to s^{-4} at higher angles following the so-called Porod's law for homogeneous particles.

1.2 SAXS studies of biological macromolecules

As the direct structure reconstruction is impossible, the main approach when interpreting SAS data utilizes the inverse approach, creating structural models such that the computed scattering matches the observed experimental pattern. In the past, this was done by trial-and-error methods; now, the modeling is done directly by

computers, which also allow one to restrict the possible solutions by *a priori* information from other methods.

SAXS can and has been used in a variety of biological problems (Fig. 1.2). If no other information is known for a macromolecule it can be used to reconstruct low resolution models *ab initio* (Esposito *et al.* 2008, Funari *et al.* 2000). However, SAXS becomes much more powerful when combined with other methods. High resolution models determined by crystallography and nuclear magnetic resonance (NMR) can be used to build models of complexes by fitting the SAXS data (granted that no major conformation change occurs). If the interface of the components of the complex is known (e.g. from mutagenesis studies, fluorescence resonance energy transfer (FRET), NMR) this information can be used to reduce the ambiguity of the models created from SAXS. Residual dipolar couplings (RDCs) from NMR can determine the relative orientation of one complex component in relation to the others (Gabel *et al.* 2008). Electron microscopy models, especially on large and highly symmetric particles, help to reduce the search volume. When a mixture of different components is found in solution (for example when the binding affinity of a complex is low and both complex molecules and individual components exist) complementary techniques like analytical ultracentrifugation (AUC) or dynamic light scattering (DLS) can be used together with SAXS to determine the amount of each component in the solution (granted that their molecular weight is different enough). SAXS is also one of the few methods available to analyze the overall structure of flexible proteins (Bernado *et al.* 2007).

Since X-rays interact mainly with the electrons they may affect the integrity of the measured sample by ionization of the solute or solvent and subsequent radiation damage due to free radicals. The radiation damage problem is often treated by adding scavengers such as dithiothreitol (DTT) or protectants like glycerol to the sample before the measurement.

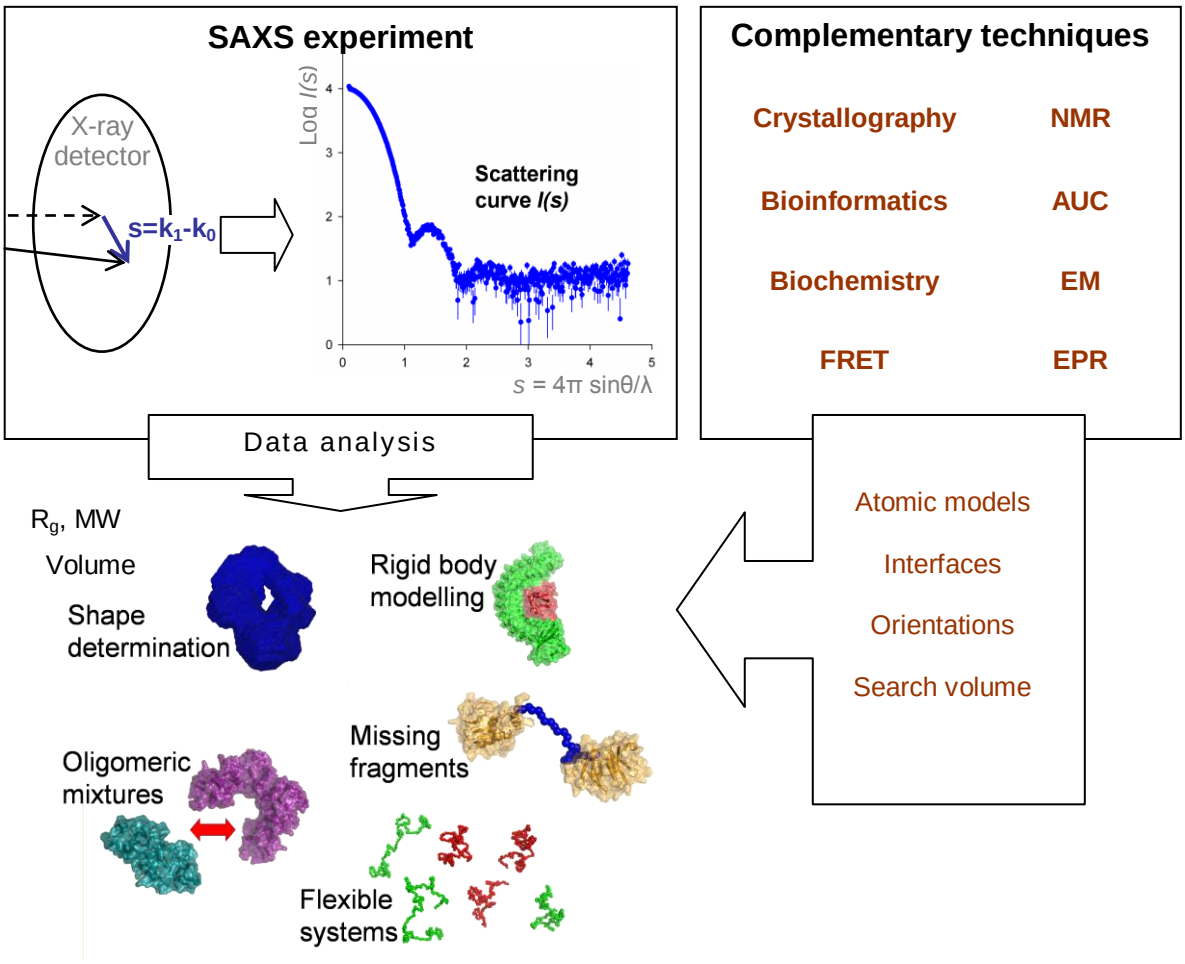


Fig. 1.2 SAXS applications in biology

In the general case the experimental scattering pattern results from a correlation of the shape of the individual molecules (form factor) and the arrangement of the molecules in the solution (structure factor). An experimental scattering pattern can thus be expressed as

$$I_{exp}(c, s) = I(s)\mathrm{SF}(c, s) \qquad \text{Equation 1.4}$$

where $I_{exp}(c, s)$ is the experimental (measured) scattering, $I(s)$ is the from factor and $\mathrm{SF}(c, s)$ is the structure factor. To construct a model for a biological macromolecule it is important to limit the effects of the structure factor to get a "pure" form factor (essentially to make $\mathrm{SF}(c, s) = 1$). At low concentrations (typically below a few mg/ml) these effects are usually small and may be neglected since the individual

molecules are not crowded enough to interact significantly with each other and cause order in the solution. Further, scattering patterns from different concentrations can be compared and, if differences are observed, extrapolated to infinite dilution.

1.3 Automation bottlenecks

The SAXS measurements produce immense amounts of data, especially on modern 3rd generation synchrotron sources. Recent theoretical developments make it possible to retrieve significantly more structural information about biological macromolecules from the high quality SAXS data than previously believed (Petoukhov & Svergun 2007). Like in high-throughput macromolecular crystallography (Winn *et al.* 2002) large-scale analysis of proteins and macromolecular complexes is emerging also in SAXS. A crucial step towards the extensive use of synchrotron sources in high-throughput crystallography was the introduction of automatic sample changers and remote operations (McPhillips *et al.* 2002).

Several points need to be considered for the developing field of high-throughput SAXS measurements and automated procedures. The first aspect that would profit from automation is the SAXS experiment itself. Synchrotron radiation experiments require repetitive measurements performed in a consistent way many times within a limited time frame dependent on the length of time allocated to the experiment. In order to reduce human errors, which are inherent in the taking of repetitive measurements, especially after working long hours, automation of the experiment is highly desirable. This is especially important when dealing with biological samples, as they are often available in limited quantities and obtained after extensive preparative work.

The second point that needs to be taken into account for high-throughput SAXS measurements is the convenient storage of large amounts of collected data along with the experimental setup details as well as obtained results. Here it is of vital importance to use common data storage formats.

On modern 3rd generation synchrotron sources precise SAXS patterns can be collected in a fraction of a second (undulator beamline) or within a few minutes (bending magnet beamline), whereas the manual changing of the sample takes longer than the

actual exposure. However, the most time consuming step that requires hours to months is the manual data processing and the construction of the structural models.

A prompt analysis of the data directly after measuring would be highly desirable, since an immediate estimate of the data quality would allow the repetition of the experiment if necessary. This would accelerate the data processing as well as data quality and would help to avoid false results. However, this would also require time consuming user intervention during the data analysis and interpretation process. Again, an ideal solution is the automation of the measurement and analysis process in combination with an adequate quality control.

Thus, automation of data analysis becomes an indispensable prerequisite for adequate evaluation of high-throughput SAXS experiments. The first solution scattering sample changer was installed at the X33 beamline of the EMBL, at the storage ring DORIS-III, DESY (Round *et al.* 2008). In the X33 sample changer the solutions are kept in thermostatically controlled well plates allowing for operation with up to 192 samples. The measuring protocol involves controlled loading of sample solutions and matching buffers, followed by cleaning and drying of the cell between measurements. A throughput of approximately 12 samples per hour, with a failure rate of sample loading of less than 0.5%, was observed for 90% of solutions, although depending on the solvent composition and/or on the sample properties the loading failure rate may increase dramatically and manual correction is required. The sample changer is controlled by a client-server-based network protocol, locally and remotely. Full integration with the beamline control software allows for automated data collection of all samples loaded into the machine with remote control from the user, though lack of flexible means of automated data processing prevents large-scale remote operation. The feedback from over 100 external user groups that used the setup during 2009 indicates that the ease of use and reliability of the user operation at the beamline were greatly improved compared with the manual filling mode.

A combined system including an automated sampler changer robot and a high-performance liquid chromatography (HPLC) device has been developed at the SWING beamline of Synchrotron SOLEIL (David & Perez 2009). It is possible either to inject a small volume of sample ready for immediate SAXS analysis or to inject the

sample through an HPLC purification pathway, without manual intervention. In the sample changer mode, a few microliters of sample can be injected between two air bubbles and circulated at a controlled speed of typically 40 µl/min. A maximum of 14 samples per hour could be measured in this mode by remote controlling the sample injections. In the HPLC mode, an initially polydisperse sample can be separated into each of its components before immediate data acquisition. The sample cell is thermostated, and offers online UV–Vis absorption monitoring. A video camera focused at the capillary position allows permanent visualization of the sample while data are collected but no automated image analysis is implemented and only visual inspection can verify successful sample loading.

Among other automated data collection approaches one can mention the microfluidic "lab-on-a-chip" (Toft *et al.* 2008) and automation of the BL4-2 beamline at SSRL, at the same time automated on-line data processing is hardly available.

A high-throughput SAXS data collection and analysis pipeline for structural characterization of proteins in solution is implemented at the SYBYLS beamline at the Advanced Light Source (Hura *et al.* 2009). A Hamilton pipetting robot is used to transfer the samples in a needle from a 96-well plate to the helium-filled sample holder, providing an anaerobic environment with low X-ray scattering cross-section and reduced background. Both sample cell and the 96-well plate are temperature-controlled. The implemented data analysis protocol employs programs from the ATSAS package (Konarev *et al.* 2006) automating to a certain extend the data flow with Perl scripts. While comparison of calculated scattering profiles to experimental data, *ab initio* shape determination and rigid body modeling are reported as fully automated, primary data analysis steps such as evaluation of R_g, D_{max} and molecular mass are done manually using PRIMUS (Konarev *et al.* 2003).

The Windows-based data processing and reduction program PRIMUS is an interactive suite with a graphical user interface (GUI) that allows to perform basic data manipulations and to compute overall structural parameters such as R_g, $I(0)$ and Porod volume. The suite provides interfaces to other interactive data manipulation software including the indirect Fourier transform program GNOM (Svergun 1992) and modeling by simple geometrical bodies.

Another example of an interactive GUI-based package for small-angle scattering data analysis is Irena (Ilavsky & Jemian 2009). It is based on the commercial *Igor Pro* application and is primarily oriented towards investigation of data from non-biological systems. Certain tools such as Guinier and Porod fits and the pair-distribution function are applicable for monodisperse solution scattering; however their usage involves multiple manual runs with slightly different optimization conditions. Although containing individual elements of automation (scripting tools for batch processing and a record-keeping system) the package is hardly suitable for large-scale automated data processing.

The aforementioned ATSAS package is primarily oriented to the analysis of solutions of biological macromolecules, but it can also be used for non-biological systems yielding one-dimensional isotropic scattering patterns. Apart from PRIMUS and some other interactive programs this package includes cross-platform command-line tools for *ab initio* low-resolution shape determination (DAMMIN (Svergun 1999), DAMMIF (Franke & Svergun 2009) and GASBOR (Svergun *et al.* 2001) for single-component particles, MONSA (Svergun & Nierhaus 2000) for multi-component particles); calculation of scattering profiles from atomic models of macromolecular structures (CRYSOL (Svergun *et al.* 1995) for X-rays and CRYSON (Svergun *et al.* 1998) for neutrons); rigid body modeling of macromolecular complexes (SASREF and BUNCH (Petoukhov & Svergun 2005)). Availability of such cross-platform console applications clears the way to automation of large-scale data analysis and, as seen above, first steps are made using primitive scripts, but there is still a gap between radial averaging and 3D-modeling. The lack of well-established experimental data quality control procedures and automated primary data reduction software is a factor limiting further progress of automated model building and remote operation.

1.4 Scope of this thesis

To make large-scale high throughput SAXS studies possible one needs to solve the problems of reliability of the experiment, automation of the data processing and control of the input and output data quality. The major steps along the SAXS experiment and data analysis are in focus of this thesis:

- experimental data collection;

- tools for automated data processing;
- automated pipeline for on-line data analysis;
- convenient storage of the obtained results.

Chapter 2 is dedicated to automation of the experiment, in particular to the control of the sample cell filling. The new visualization hardware setup and the image analysis software module not only make unattended measurements possible but also allow remote operation/supervision.

In chapter 3 hardware-independent data reduction and analysis tools are presented. A novel tool for automated estimation of such fundamental parameters as radius of gyration, forward scattering and experimental data quality clears the way for comprehensive automated characterization of the sample. A set of data manipulation tools allows performing all basic data reduction steps including extrapolation to zero solute concentration.

An automated data analysis pipeline that covers major processing and interpretation steps from primary data reduction to low-resolution three-dimensional model building is presented in chapter 4. The pipeline integrates the tools from the preceding chapter with legacy software modules; a novel flexible way of controlling data processing steps with cross-validation of intermediate results ensures persistent data analysis without user intervention. A user-friendly web-compatible XML format was developed to store the obtained results.

Automated sample loading and the data processing pipeline were extensively used by a number of external user groups. Chapter 5 describes three user projects that benefited from the results of this work.

Although this thesis was focused on biological small angle X-ray scattering the methods described in chapters 3 and 4 are also applicable to small angle neutron scattering (SANS) data and, to some extent, to non-biological samples in solution.

2 Automation of the SAXS experiment

In a typical SAXS experiment a small volume of sample solution is injected into a cell where it is exposed to an intense synchrotron X-ray beam and the scattering pattern around the primary beam is recorded by a 2D detector that is placed at a certain distance behind the sample cell (Fig. 2.1). Automation of this procedure is critical for a high-throughput facility and became possible due to the recent advancements in robotic liquid handling. Proper filling of the sample cell is a necessary condition for a solution scattering experiment. An automated sample filling control system was implemented at the X33 SAXS beamline of the Hamburg Outstation of the European Molecular Biology Laboratory (EMBL) and is described in this chapter.

2.1 X33 beamline setup

The bending magnet SAXS beamline X33 of the EMBL at the DESY DORIS-III storage ring is dedicated to solution X-ray scattering studies of biomacromolecules (Roessle *et al.* 2007). The synchrotron beam is monochromatized and focused horizontally using a bendable germanium (111) crystal located at 21m distance from a source. A gravimetrically bent rhodium coated Zerodur mirror mounted after the monochromator serves for suppression of higher harmonics and vertical focusing. The monochromatic beam is shaped by a pair of slit systems; the incident beam intensity monitor and the beam shutter are placed after the second slit system (Fig. 2.1). The sample cell is located inside a vacuum chamber at a distance of 28.5 m from the source; 2.7 m downstream from the cell the Pilatus 1M detector is placed. The cell temperature is adjusted using an external Huber Ministat water bath; a temperature feedback loop uses the signal from PT100 temperature sensors attached to the cell. The sample solution (typically at a concentration between 1 and 10 mg/ml) is transferred to the cell through a plastic tubing by an automated sample changer (Round *et al.* 2008) that performs sequential loading of samples and cleans the tube path. The loading process is controlled by a standard Baumer FVDK 10P69Y0 flow sensor installed inside the sample changer. The capacity of the sample changer is 2x96 Eppendorf tubes of 200µl volume for samples and 2x24 tubes of 1.5ml volume

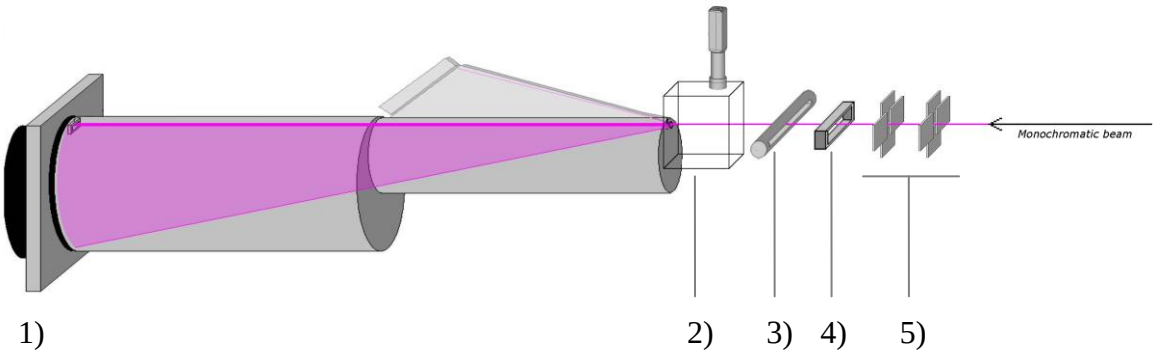


Fig. 2.1 Components of the X33 beamline: (1) SAXS detector; (2) vacuum chamber with the sample cell connected to the automated sample changer; (3) experimental shutter; (4) incident beam monitor; (5) slit systems.

for buffers. In case of specific samples (high viscosity or amounts less than 50 µl) manual filling by a syringe can optionally be used.

2.2 Control of the sample cell filling

Depending on the physical properties of samples and/or corresponding buffers (density, viscosity, detergent content etc.) two kinds of problems may occur during the sample loading process: a) the flow sensor does not recognize the liquid being filled and the sample changer fails to load the specimen, b) bubbles are formed in the cell (Fig. 2.3, right). These problems make useful data collection impossible; therefore to control such sample loading issues and achieve fully automated sample loading it is crucial to provide visual feedback. The sample cell is visualized by a video camera mounted on the top plate of the vacuum chamber (Fig. 2.4), and a mirror positioned at an angle of 45° to the horizontal axis of the cell (Fig. 2.2 2). The mirror is fixed on a shaft inserted in the vacuum feedthrough mounted on the back side of the chamber. The shaft is moved by the pneumatic rotary actuator by 90º synchronized with the experimental shutter, such that during exposure the mirror moves out of the beam path. Improper cell filling can be corrected by moving the solute back and forth by activating the corresponding sample changer pump.

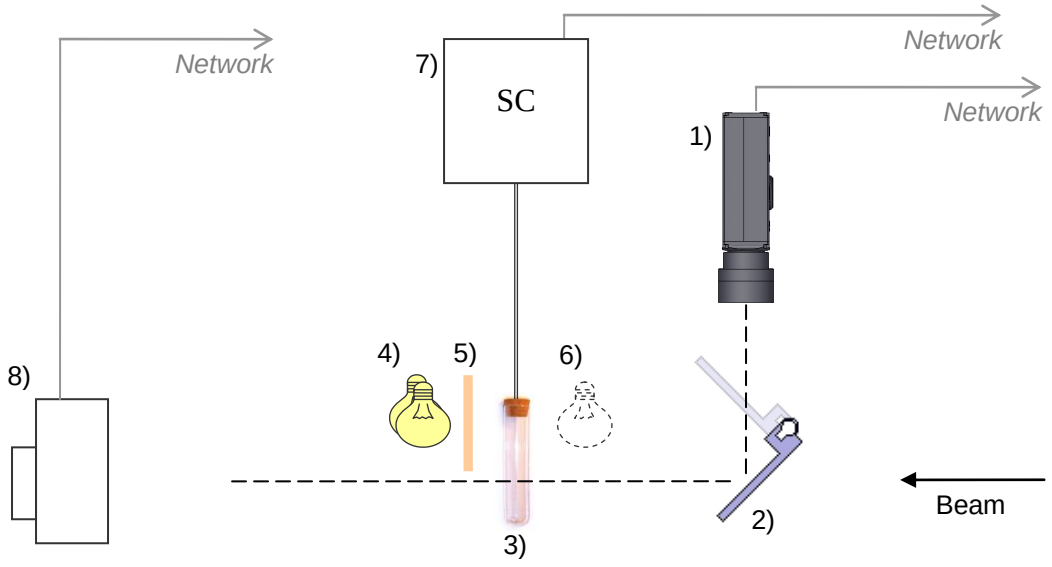


Fig. 2.2 Visualization setup: (1) camera; (2) movable mirror ; (3) sample cell; (4) six LEDs; (5) translucent screen; (6) old position of the LED; (7) sample changer; (8) X-ray detector. During the exposure the mirror is moved to the upper position in order not to block the beam.

Previously a simple analog video camera was used for visual inspection of the sample cell. It gave the users the opportunity to monitor the sample filling process and to correct eventual problems manually. However, to allow for unattended measurements automated control of the filling is required. The first approach to achieve the automation involved digitizing the images of the analog camera using a National Instruments frame grabber which provided 400x180px pictures of the region of interest. Initial tests showed that the image quality was not sufficient to ensure reliable automated image analysis due to low contrast, a high noise level and unsatisfactory lighting conditions in the cell (see Fig. 2.3 a). Even by visual inspection it was difficult to make a decision whether the cell is properly filled.

For the new cell setup a high resolution (3 megapixel) digital network camera Elphel NC353L was chosen. This camera is able to provide 15-30 frames per second to multiple clients over an Ethernet connection. The shutter speed (exposure time) can be adjusted in a broad range from milliseconds to a few seconds. The ability to deliver JPEG images over the HTTP protocol makes it easy to provide web-based remote visual inspection. With the lens combination of MMS OBJ-11 and MMS R-3 manufactured by Edmund Optics it is possible to take images of the region of interest of 1500x680px size. The sample cell is illuminated from the back side of the container by six 5mm focusing casing LEDs (44cd each), three from the right side and three from the left side of the cell. An improved contrast is achieved by a setup that avoids direct light from the LEDs on the lens of the camera. We introduced a translucent screen to reach diffuse illumination and to avoid unwanted deflections. Optionally, the normal white LEDs can be replaced by red light LEDs which is convenient for measuring light-sensitive samples. The latter setup was successfully used for the measurements of YtvA photoreceptor protein from *Bacillus subtilis* (see section 5.2).

Since a properly filled cell is visually indistinguishable from an empty cell, the cell filling is controlled by monitoring the differences between successive frames that are taken by the camera during the filling process. The frames are compared two by two, and if the difference is higher than a certain adjustable threshold, a conclusion that the sample has reached the cell is made. The new setup also makes it possible to reliably detect bubbles from a single image: after the cell is filled, the current snapshot is

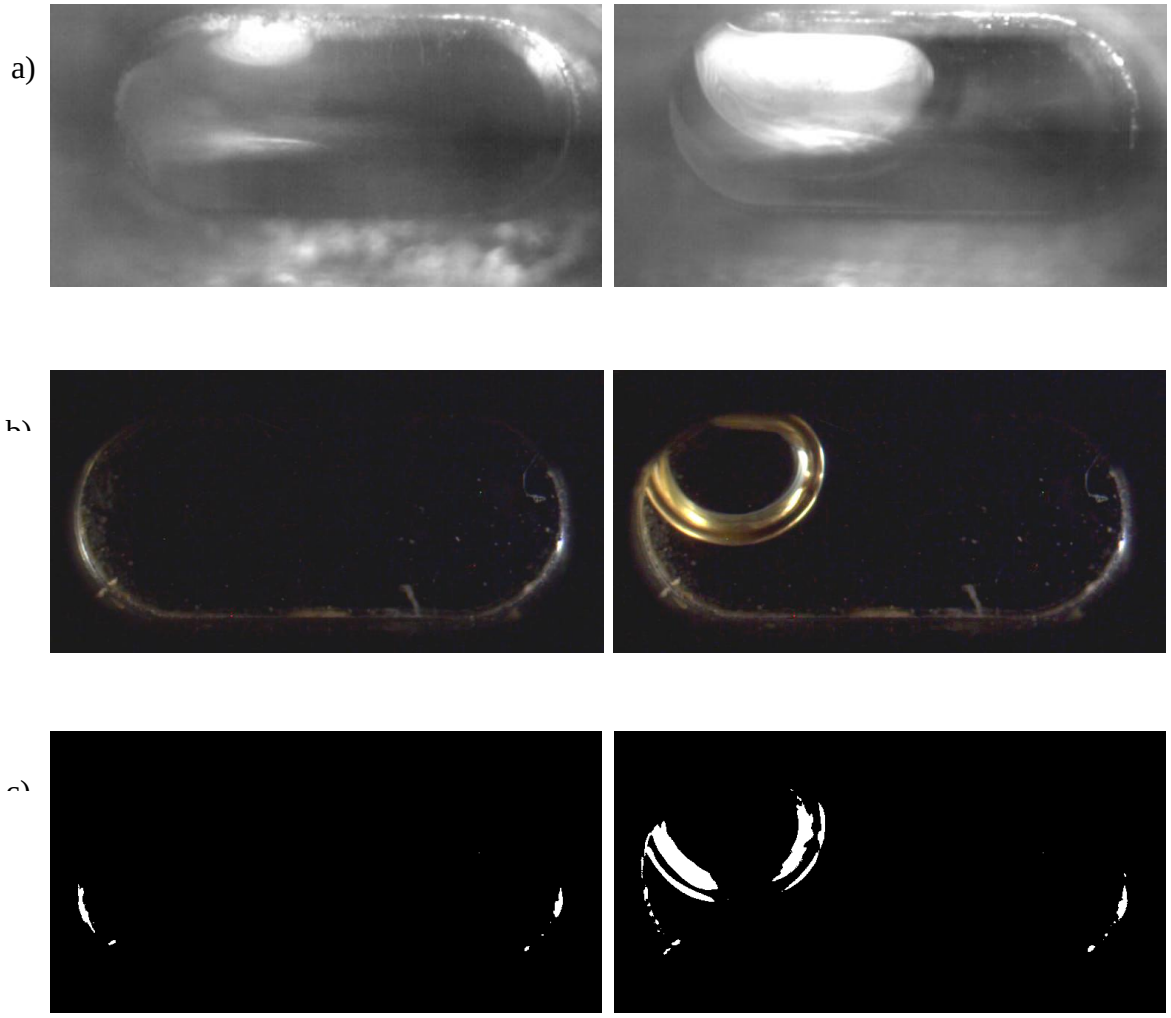


Fig. 2.3 Images from the sample cell camera. Left: properly filled cell, right: filled with a bubble. a) old setup, 8 successive frames were averaged to reduce noise but the bubble is hard to detect; b) new setup; c) the new setup images as seen by the brightness threshold bubble detection algorithm.

checked for areas of high brightness (Fig. 2.3 c). Such areas are automatically registered as bubbles. The brightness threshold and the size of the areas can be adjusted using a calibration algorithm. Tests have shown that such algorithm is more reliable than a contour-tracing approach. If improper cell filling is recognized, the system attempts to correct the filling; alternatively the system can alert the user giving him an opportunity to adjust the filling by manually controlling the sample changer pumps through a simple graphical interface. For logging purposes snapshots of the cell are saved before and after every measurement. An additional digital network camera is used for the surveillance of the sample changer during attended or remote operations. The camera is identical to the one used to control the sample filling process and thus can be used as a backup.

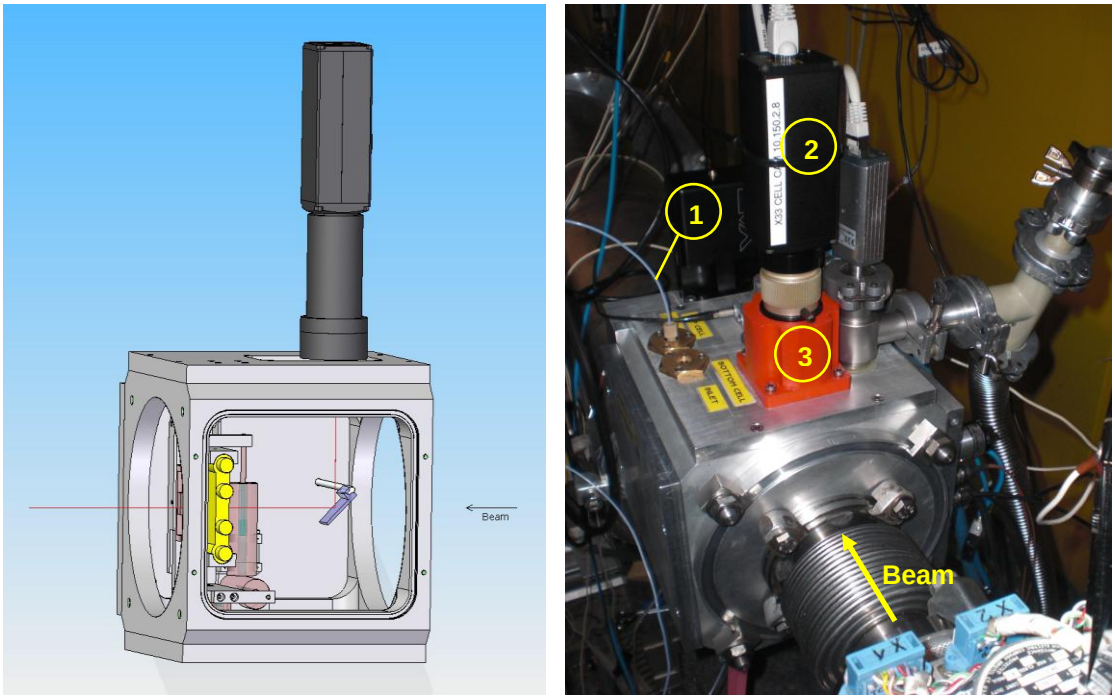


Fig. 2.4 The vacuum sample cell chamber at the X33 beamline. Left: workshop drawing; right: assembled, view from above: (1) sample inlet; (2) digital camera; (3) lens casing.

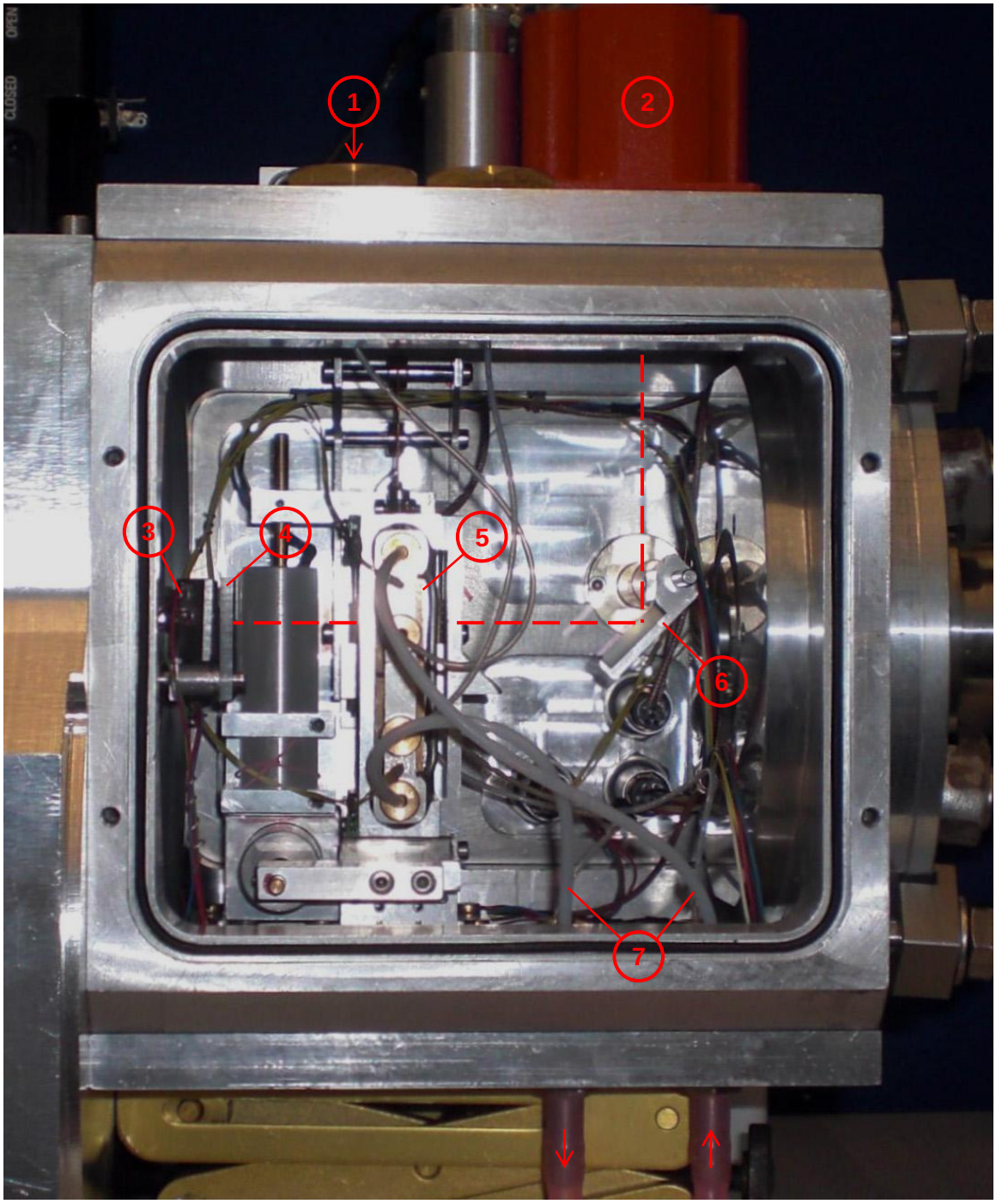


Fig. 2.5 The vacuum sample cell chamber with removed front cover: (1) sample inlet; (2) lens casing; (3) LEDs; (4) translucent screen; (5) sample cell; (6) movable mirror; (7) water bath circuit.

2.3 Software implementation details

To achieve full automation of the X33 beamline operation we developed a general approach for high throughput SAXS data collection. The Beamline Meta Server (BMS) links automated sample loading and primary data reduction, allowing for queuing of multiple measurements for subsequent execution and providing means of remote experiment control. The BMS is integrated into the existing TINE control-system (Bartkiewicz & Duval 2007) of the DORIS-III storage ring at DESY. Conceptually, it is an abstract and hardware-independent representation of the entire beamline that communicates with hardware abstraction TINE-servers which were implemented for existing hardware devices. These include the detector, the cell temperature control, the motor control subsystem to adjust the slits and other movable parts of the beamline, the sample changer and the cell filling control module. To close the gap between hardware automation and automated data reduction and processing, the BMS also supports "software devices", namely a program for radial averaging of the 2D data and appropriate normalization and an automatic logbook of monitor values such as cell temperature and current beam characteristics.

To achieve real-time operation, the cell filling control server includes a class that processes the images in parallel to the communication with the BMS and the sample changer server. This class sends an HTTP request to the camera and waits for reply in a separate thread. As soon as the next image is downloaded, the class decompresses the JPEG format and performs appropriate image processing. Same approach is used for visualizing the cell in the BMS graphical user interface.

2.4 Results

Implementation of the new design of the sample visualization setup coupled with an image recognition module provides sensitive control over sample loading and dramatically reduces the number of failures during the measurement. The sample loading control hardware and software has been developed as a key component for the automated data collection setup. Full integration with the beamline control software and coupling with the automated data analysis pipeline permits fully automated and remote controlled high-throughput SAXS studies at the X33 beamline.

3 Tools for automated data processing and analysis

Processing the data obtained from an automated SAXS experiment is the next essential step both for unattended measurements and user-controlled experiments. Automated characterization of the specimen minimizes possible human errors ensuring reliable and reproducible outcome. Obtaining immediate feedback about the sample quality and its overall parameters allows one to correct the sample preparation conditions if necessary. Hardware-independent tools for automated primary data reduction and analysis are described in this chapter.

3.1 Data reduction and analysis steps

The elastic scattering of randomly oriented particles in solution and the background due to solvent scattering (and other contributions) result in an isotropic (1D) pattern. This pattern is usually recorded on a 2D detector to improve the counting statistics of the scattering data. In order to normalize the recorded pattern and put the measurements on an absolute scale, the intensity of the transmitted direct beam is measured using different methods described in (Koch, M.H. *et al.* 2003). After scaling against the transmitted beam intensity and exposure time the normalized 2D scattering pattern is transformed into 1D arrays of scattering intensities $I(s)$ and their associated errors as a function of the modulus of the scattering vector s. These steps depend on the detector type and the beamline design and are therefore normally performed directly at the beamline right after the measurement, either manually or automatically.

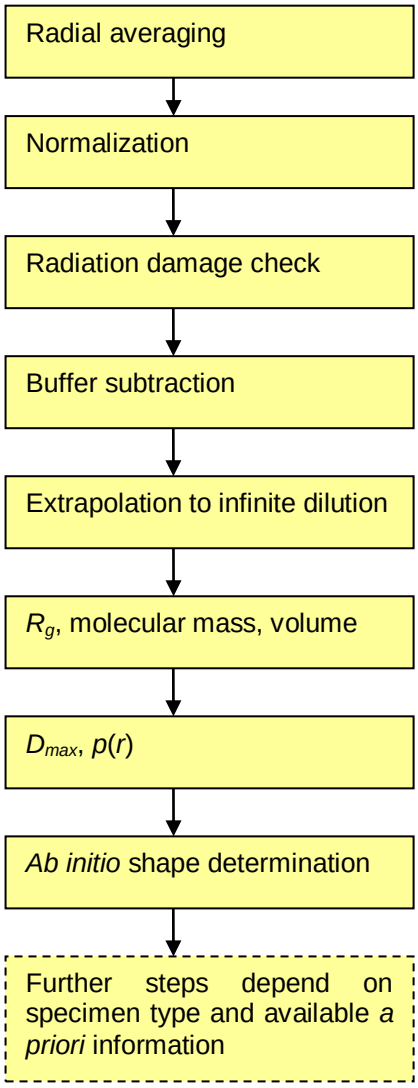


Fig. 3.1 Typical SAXS data analysis steps

The scattering pattern of the macromolecular solute is obtained by subtracting the scattering of the buffer at dialysis equilibrium that is measured separately in addition to the macromolecule solution. The subtraction of the buffer must be done very accurately to correctly determine the difference between the experimental patterns of the solution and of the solvent. The difference is usually very small, both at very small angles, where the signal is dominated by the background due to the primary beam, and at large angles, where the intensity is largely due to the scattering from the solvent. Minor movements of the incident beam during the experiment may lead to instabilities of the background subtraction; therefore the measurement of the solute is typically surrounded by two buffer measurements. The scattering

patterns of the sample and the buffer are normally measured in the same sample cell to avoid subtracting the contribution of the empty cell. To monitor possible radiation damages to the macromolecule under investigation two or more exposures of the same sample are recorded. The concentration of the macromolecular solute must be accurately determined to achieve correct normalization against this parameter, therefore absorbance measurements at 280 nm are normally carried out directly before the SAXS measurement.

Interparticle interference may influence the initial part of the scattering curve. To eliminate the influence one usually extrapolates to infinite dilution, namely, having measured several samples with different concentrations one reduces the data to the "zero" concentration situation assuming linear dependence of interference effects on concentration. This extrapolation works well for repulsive interactions, but may be more difficult to apply in the case of attractive interactions, leading to oligomerisation or aggregation. The scattering from aggregates may influence the entire dataset making further data processing impossible, therefore if aggregation is detected one should attempt to remove it by centrifugation, filtration or varying buffer conditions.

Several overall parameters (invariants) can be evaluated directly from the SAXS patterns of sufficient quality: molecular mass, radius of gyration (R_g) and hydrated volume. The shape of the distance distribution function $p(r)$ provides information about the main features of the shape and size of the solute particles, including the maximum particle diameter (D_{max}).

The experimental scattering curve can further be analyzed by *ab initio* modeling methods, which yield a low resolution shape of the molecule under investigation (Svergun 1999). Many models obtained this way can be subsequently compared and averaged to determine common structural features (Volkov & Svergun 2003). Further modeling steps depend on information available from complementary methods.

It is important to evaluate the data immediately after the measurement to correct the experimental conditions if necessary. Data evaluation is normally performed manually by the user but the lack of experience or human errors may lead to generating incorrect results at any data analysis step. Automation of evaluation of radius of gyration and other overall parameters help to overcome these difficulties.

3.2 Automated estimation of radius of gyration

There are two essential parameters obtained directly from the scattering profile of a biomolecule in solution: the radius of gyration (R_g) and the forward scattering intensity ($I(0)$).

The radius of gyration is the average of square center-of-mass distances in the molecule weighted by the scattering length density. R_g is an important measure for the overall size of a macromolecule, its automated estimation is essential for further automation of data analysis. Observing changes of R_g allows monitoring of processes like folding/unfolding or oligomerization of the investigated protein under varying physicochemical conditions, e.g. temperature, concentration, solvent composition.

Routinely the radius of gyration is estimated either from the low angles part of the curve using the Guinier approximation as described below or from the $p(r)$ function essentially using the whole scattering curve. Obviously the two R_g values obtained from these two methods should be close, i.e. knowing the R_g value from the Guinier approximation allows us to automate the calculation of the distance distribution function.

3.2.1 The Guinier approximation

In Equation 1.1 we can use the Mclaurin series

$$\sin(sr)/sr = 1 - s^2r^2/3! + s^4r^4/5! - \ldots$$

If we restrict ourselves to the first two terms, then in the vicinity of $s = 0$

$$I(s) = 4\pi \int_0^{D_{\max}} p(r)\frac{\sin(sr)}{sr}dr \approx 4\pi \int_0^{D_{\max}} p(r)(1-\frac{s^2r^2}{6})dr$$

Substituting $I(0) = 4\pi \int_0^{D_{\max}} p(r)dr$ and $R_g^2 = \frac{\int_0^{D_{\max}} p(r)r^2dr}{2\int_0^{D_{\max}} p(r)dr}$ we get

$$I(s) = I(0)(1 - s^2R_g^2/3).$$

Using the Mclaurin series $\exp(-s^2R_g^2/3) = 1 - R_g^2s^2/3 + O(s^4)$ we can write

$$I(s) = I(0)\exp(-s^2R_g^2/3). \qquad \text{Equation 3.1}$$

This is the well known Guinier approximation (Guinier 1939) which is valid for sufficiently small scattering vectors (in the range up to $sR_g \leq 1.3$). The value of R_g is estimated from the slope of the linear fit of Ln[$I(s)$] *versus* s^2 (a Guinier plot), it's intercept gives $I(0)$, see Fig. 3.2. Linearity of the Guinier plot is a sensitive indicator of the quality of the experimental data, and deviations from linearity usually point to strong interference effects, polydispersity of the samples or improper background subtraction. Despite the simplicity of the Guinier formula, automated computation of R_g is not a trivial task, in particular because of uncertainty in the fitting interval. Visual inspection is most often used to select the range of the Guinier fit, and this interactive fitting can be conveniently done in several packages including *PRIMUS* (Konarev *et al.* 2003).

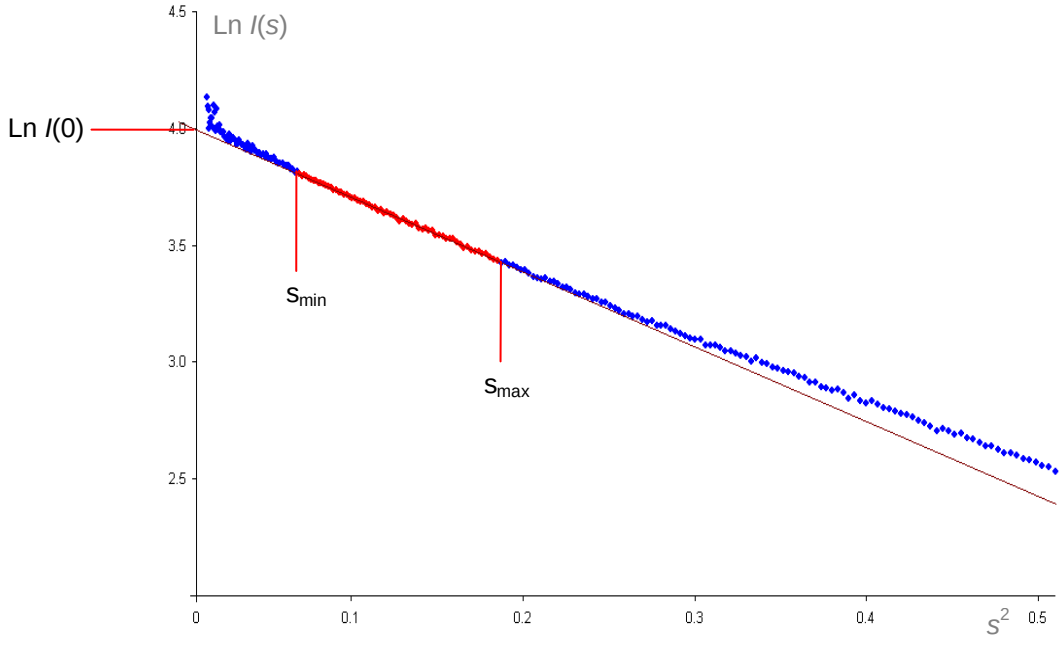


Fig. 3.2 The Guinier plot for a 8 mg/ml BSA solution scattering data: ln[$I(s)$] *versus* s^2. s_{min} to s_{max} is the so-called Guinier region (red); $s_{min}R_g < 1$, $s_{max}R_g \leq 1.3$.

3.2.2 Molecular mass estimation

The value of $I(0)$ obtained after scaling for concentration corresponds to the scattering of a single particle and is proportional to the square of the total excess scattering length in the particle (Koch, M.H.J. *et al.* 2003). If the measurements are made on an absolute scale, $I(0)$ can be directly related to the molecular mass of the solute:

$$M = I(0)\frac{\mu^2}{(1-\rho_0\psi)^2 N_A}$$

For proteins the ratio between the molecular mass and the number of electrons in the particle, μ, has a value close to 1.87. ρ_0 is the average electron density of the solvent (in e/nm^3), ψ the ratio of the volume of the particle to its number of electrons, N_A is Avogadro's number. In fact the intensity values are normally obtained on a relative scale which is dependent on the X-ray detector but it is possible to calculate the absolute values by scaling against scattering of a known reference, e.g. water.

For SAXS with solutions of biological macromolecules, it is often easier to measure a fresh solution of a well-characterized protein (e.g. bovine serum albumin or lysozyme) as a reference, assuming that its partial specific volume is similar to the one of the samples. It should be measured in the same conditions as the samples, with accurately determined concentrations. In this case the molecular mass of the solute is calculated as:

$$M = \frac{I(0)M_{ref}}{I_{ref}(0)} \qquad \text{Equation 3.2}$$

where M_{ref} and $I_{ref}(0)$ are the known molecular mass and measured forward scattering intensity of the reference protein respectively. Here we assume that the electron density of the sample and the reference protein is similar. Since the density corresponds to the hydrated molecule, i.e. the protein with its hydration shell surrounding it (which has different density than the bulk solvent), this assumption is not perfectly correct since not all proteins are folded and hydrated to the same extent and the surface to volume ratio of molecules of different size can vary (Mylonas &

Svergun 2007). Obviously nucleic acids have significantly larger density than proteins because of high content of heavy phosphorus atoms.

The molecular mass computed from $I(0)$ is required for the qualitative appraisal of the SAXS data but it is also an important result in itself providing information on the oligomeric state of the molecule in semi-physiological conditions. The molecular mass can also be estimated from the particle volume using two other independent methods: the Porod volume from Equation 1.3 and the volume of an *ab initio* dummy atom model. The comparison of the molecular masses from these three different sources provides us with a convenient quality control of the automated data processing results.

3.2.3 The AUTORG tool

Despite the importance of the concept of the radius of gyration for SAXS, publicly available programs for automated R_g determination do not seem to be available. Since 1939, when A. Guinier has first introduced this parameter, scientists were using various interactive procedures to select the fitting range for R_g in [Ln $I(s)$, s^2] coordinates. We have developed a program *AUTORG* for a fully automated determination of R_g from the scattering data (Petoukhov *et al.*, 2007). The program also estimates the quality of the fit and provides information for other modules of the automated pipeline. Three versions of the program are publicly available: a menu-driven Windows application with a simple graphical user interface (GUI), a Windows dynamic link library (DLL) and a cross-platform console application with command line input/output; the latter is incorporated in the automated data analysis pipeline. The program is written in C++.

Like all modularized tools described below, *AUTORG* uses the *libsaxsdocument* library to read experimental data files. This library was developed to provide a format-independent interface for reading and writing data files in various ASCII formats. The modular design of the library potentially allows one to incorporate existing data formats such as SasCIF (Malfois & Svergun 2000), NEXUS (Maddison *et al.* 1997) or XML-based formats without changing each particular tool every time a new format is added. The data file in plain ASCII format is expected to contain three

columns: (1) momentum transfer, (2) experimental intensity and (3) experimental errors. If errors are not present, they are estimated as 4% of the intensity.

The GUI version can be started from another application (e.g. PRIMUS), from the Windows Explorer context menu or it can be used as a stand-alone application. After opening the data file AUTORG displays the R_g value and its standard deviation, the $I(0)$ value, range of points used as the Guinier interval, sR_g limits and the data quality

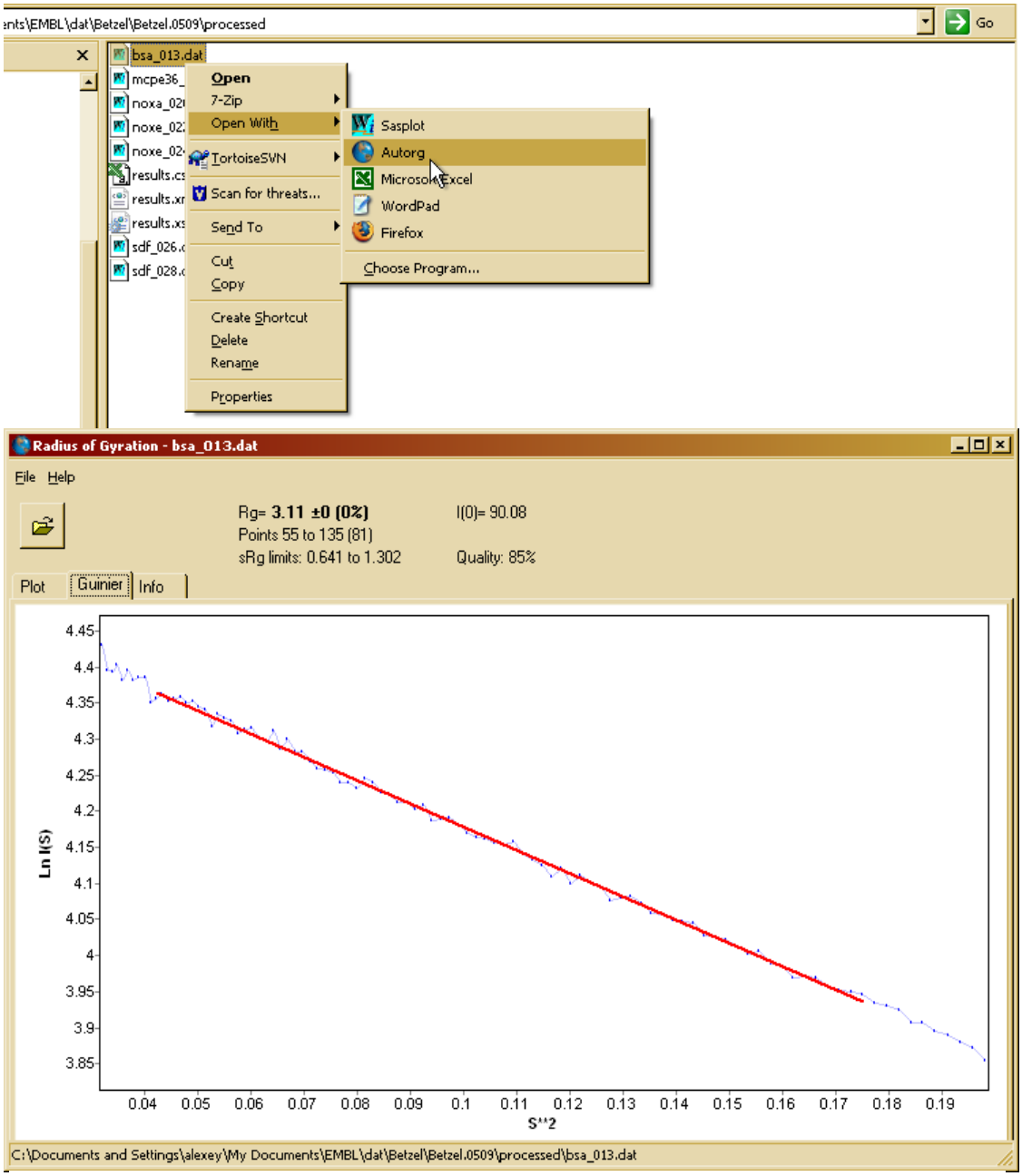


Fig. 3.3 AUTORG can be accessed from the context menu (above) and has a simple Windows GUI (below)

estimate. Besides, the application also plots the best obtained Guinier fit for visual inspection (Fig. 3.3, red line). A [Log *I*(*s*), *s*] plot is available under the "Plot" tab; all additional information that could be read from the file (besides intensities) is displayed under the "Info" tab.

The console version is used as follows:

```
$> autorg [OPTIONS] <DATAFILE(S)>
```

where the options are:

- -o, --output <FILENAME>: relative or absolute path to save result(s). If the output file name has a ".csv" extension the output will be in comma separated values (CSV) format unless the output format is specified (see next option).
- -f, --format <FORMAT>: output format, one of: 'csv', 'ssv', 'table'.
    - 'csv' will force to produce the output in Excel-compatible comma separated values format, a header will be written.
    - 'ssv' will force to produce the output in machine-readable space separated format without a header; the sequence of values is same as in CSV format apart from the file name – it will be written last, not first. This format is used by the automated data analysis pipeline.
    - 'table' will force to produce the output as a human-readable table with rounded values. This is the default value if more than one input file is given.
- --mininterval <NUMBER>: minimum acceptable Guinier interval length in points. Default is '10'.
- --smaxrg <NUMBER>: maximum acceptable $s_{max}R_g$ value. Default is '1.3' which works well for globular proteins. However, one might use values closer to 0.8 for elongated particles, multidomain proteins with long linkers, or other extended semi-flexible macromolecules (Putnam *et al.* 2007).
- --sminrg <NUMBER>: maximum acceptable $s_{min}R_g$ value. Default is '1.0'.

- -v; --version: print version information and exit.
- -h, --help: print a summary of arguments and options and exit.

If the '--output' option is not specified the output will be written to STDOUT and no output files will be produced. The output includes:

- DATAFILE name.
- Estimated R_g; if the input data was in $Å^{-1}$ the unit of R_g will be Å, if the input data was in nm^{-1} the unit of R_g will be nm.
- Accuracy of the R_g value estimated by taking into account not only the standard deviation of the selected fit as usual but also the deviation of R_g values calculated from other consistent intervals, accounting to some extent for systematic errors in the R_g determination.
- Extrapolated scattering intensity at zero angle $I(0)$ (forward scattering).
- Standard deviation of the $I(0)$ value.
- First point of the Guinier interval. Counting of data points starts from 1. Data points before the first point should not be used in further data processing because they do not follow the Guinier law.
- Last point of the Guinier interval. In table format total number of points is given in brackets.
- Estimated data quality. '1.0' means ideal quality, '0.0' – unusable data. In 'table' format it is given in percent (100% – ideal quality, 0% – unusable data). This estimation is based only on the Guinier interval (lowest angles).
- "Aggregated" flag: if aggregation was detected from the slope of the data curve at low angles the value is '1', otherwise '0'.

3.2.3.1 AUTORG algorithm

First, the input data [$I(s)$, s] is converted into the "Guinier scale" [Ln $I(s)$, s^2]. Negative intensity values that can occur after improper buffer subtraction are ignored and the ratio of negative values is calculated to be used in the data quality estimate later. Then the program selects the data range suitable for the Guinier approximation. For this, the whole range is divided into intervals divisible by given minimum interval length ('mininterval') and in each such interval (s_{min}, s_{max}) a weighted linear fit is calculated by least squares, the R_g is computed and the $s_{max}R_g$ value is checked. Intervals where $s_{max}R_g$ is less than the given maximum acceptable value ('smaxrg') are marked as "valid". Note that even in the high angles region one can find a "valid" interval with a sufficient small R_g that will satisfy the formal boundary conditions but such interval will have nothing to do with the actual Guinier approximation. Therefore the suitable data range is chosen as closest to the origin union of continuous "valid" intervals.

The search of *all possible* intervals for Guinier plots starts in the selected range: for each interval longer than a given minimum a weighted linear fit is calculated by least squares and R_g is computed. For each interval (s_{min}, s_{max}), the conditions $s_{min}R_g <$ 'sminrg' and $s_{max}R_g <$ 'smaxrg' are checked and the absence of systematic variations is verified, in which case the interval is considered consistent. Each consistent interval is rated according to its length (number of points fitted) and discrepancy (root-mean-square deviation of the fit), and the interval with the best rating is selected. The accuracy of R_g is estimated by taking into account not only the error propagation in the selected fit as usual but also the deviation of R_g values calculated from other consistent intervals (weighted by their rating), accounting to some extent for systematic errors in the R_g determination. The intercept of the fit gives ln[$I(0)$], and the accuracy of $I(0)$ is estimated in a similar way.

Thus *AUTORG* translates the perceptual criteria used during interactive R_g analysis by Guinier approximation into an algorithm to compute R_g and to estimate the quality of the fit. The program has several tunable parameters that can be adjusted by the user, such as the minimum interval length in points, the worst acceptable $s_{min}R_g$ and $s_{max}R_g$ limits, and the length and discrepancy weights used for the interval rating (not shown

in the "options" list). The default parameters are tuned to provide the most stable results and in most cases the program works automatically without any need for adjustments.

3.2.3.2 Data quality estimation

Assessment of the input data quality is crucial for automated data processing where one should avoid producing pseudo-consistent results from worthless data. Visual inspection of the SAXS curve allows an experienced user to estimate the quality by analyzing the curvature of the low-angle part. One of the characteristics derived this way is indication of strong aggregation seen as concavity of the initial part of a featureless curve. Independently of the Guinier analysis, *AUTORG* detects aggregation in a similar way. For this, the data range where the scattering intensity decays by an order of magnitude is taken, a parabola is drawn in this range using a logarithmic scale of intensity and the sign of the leading coefficient is analyzed. One should mention that lack of features and a similar concavity can also be observed with unfolded samples. To distinguish between unfolded and aggregated samples one may check the Kratky plot [$s^2I(s)$ as a function of s].

As mentioned above, linearity of the Guinier plot can be used as an indicator of the data quality. Some partially aggregated samples may show nonlinearity only over a small lowest angle region, which is discarded by the algorithm. Therefore an estimate of the overall data quality is expressed by taking into account several criteria: (a) how many consistent intervals were found, (b) how accurate is the value of R_g, (c) how many starting points were discarded, (d) how many data points with negative intensity are present, (e) whether there is indication of strong aggregation. The obtained overall estimate is made available to other programs used for automated data processing, in particular for selecting the optimum subtraction of the background. If no consistent intervals are found the search may be repeated by running the program with weakened sR_g conditions reducing the estimate of the data quality respectively.

The console version of *AUTORG* was tested on numerous data sets (see *Practical applications*) and the results were compared with those of manual R_g determination; in the vast majority of cases the automated system yielded the same results or better than those obtained by an experienced user. Estimation of the data quality separates

suspicious data to prevent further analysis and model building which could lead to false results and unjustified conclusions.

3.3 Automated extrapolation to infinite dilution

Solutions of biological macromolecules are rarely ideal and attractive or repulsive interactions influence the time-averaged spatial distribution of the particles leading to distortions of the scattering pattern (Koch, M.H.J. *et al.* 2003). If the particles are spherical (or rather globular, such that they can be considered spherical on the scale of their average separation) the measured scattering can be described by Equation 1.4 where SF(*c*, *s*) is the concentration dependent structure factor of the solution that takes into account the interactions, attractive or repulsive, between solute particles. This equation has been shown to be valid for proteins although in a restricted s range, in the case of globular particles and moderate interactions (Tardieu 1994, Vérétout *et al.* 1989). At infinite dilution SF(0, *s*) = 1, thus scattering patterns collected at several concentrations should be extrapolated to zero concentration to obtain an undistorted pattern in the low angle region. Alternatively the interaction effects can be neglected for a sufficiently small concentration and its low angle region can be merged with the higher region ($s > 2\ nm^{-1}$) of concentrated solutions ($c > 10$ mg/ml) which are generally used to obtain a sufficiently high signal-to-background ratio, especially for larger proteins, where the scattering decays rapidly at increasing *s* values. These are valid procedures because the effect of repulsive interactions is, in contrast with that of aggregation, negligible at higher angles. Attractive interactions, which often lead to unspecific aggregation, are usually characterized by a steep increase of the scattering curve at low angles as described above.

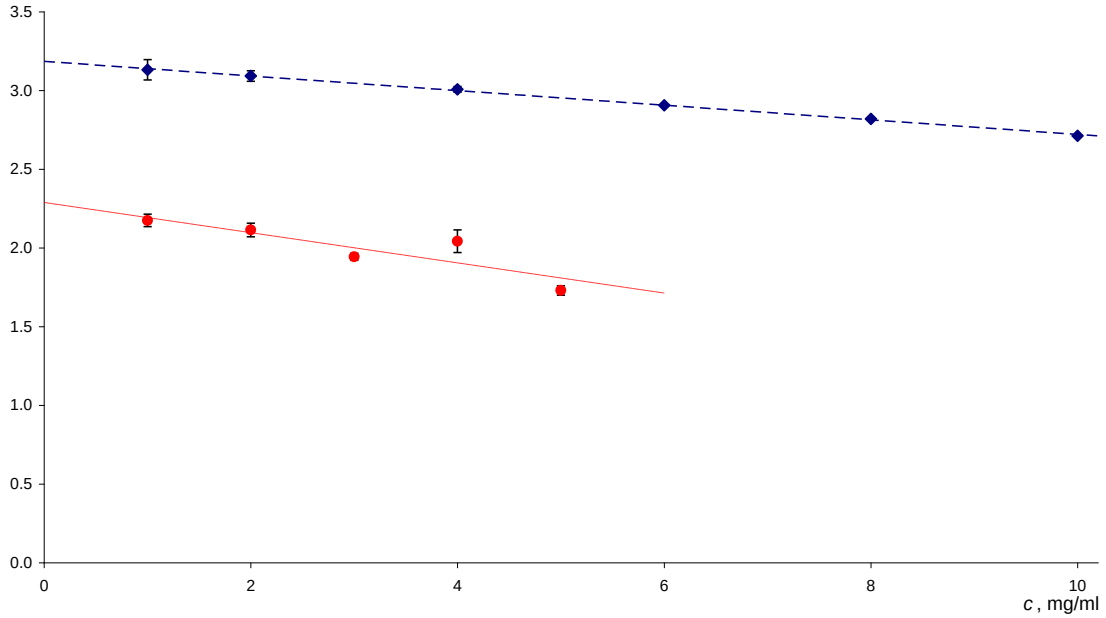


Fig. 3.4 The effect of repulsive interactions observed as linear decrease of the R_g with concentration of the non-activated full length YtvA protein (blue diamonds) and its LOV domain (red circles).

Means of interactive merging and extrapolation to infinite dilution can be found in several data reduction packages including PRIMUS. To merge two data sets from two different concentrations the user needs to manually define the range (s_1, s_2) where the data demonstrate similar behavior, assuming that the higher concentration data is distorted by particle interactions in the range (s_{min}, s_1) and the lower concentration data has unsatisfactory signal-to-noise ratio in the range (s_2, s_{max}). One of the curves is scaled in order to compensate for inaccuracy in concentration determination; the resulting composite curve consists of the range (s_{min}, s_1) of the lower concentration data, the average of both data in the range (s_1, s_2) and the range (s_2, s_{max}) of the higher concentration data. In case of reduction to "zero" concentration situation the range (s_{min}, s_1) should be extrapolated assuming linear dependence of interference effects on concentration. Absence of automation of these procedures impedes both the manual and automated data processing. Thus we have developed a program ALMERGE for automated merging and extrapolation to infinite dilution of the scattering data.

3.3.1 The ALMERGE tool

The program is designed as a cross-platform console application and is used as follows:

```
$> almerge [OPTIONS] <DATAFILE> <DATAFILE>
```

where <DATAFILE> contains experimental intensities that can be read by the *libsaxsdocument* library and the options are:

- -o, --output <FILENAME>: relative or absolute path to save the result.
- --overlap <NUMBER>: minimum overlap range length in points. Default is '50'.
- --step <NUMBER>: enumeration step in points. Default is '5'.
- -z; --zerconc: perform extrapolation to zero concentration. If this option is not specified and the output file name is present then the program will merge the two input data files.
- -c; -- concentration <NUMBER> <DATAFILE>: concentration of the following data set, overrides concentration read from the data file header. If this option is not specified and the concentration value could not be read then the data is treated as if it is not scaled against concentration.
- -v; --version: print version information and exit.
- -h, --help: print a summary of arguments and options and exit.

If the '--output' option is not specified no output files will be produced. Following information is written to stdout in a machine/human readable format: the input file names, the overlapping range in points, the range length in points, the scaling coefficient.

3.3.1.1 ALMERGE algorithm

To find the best overlap of data sets from same sample measured at different concentrations ALMERGE utilizes the algorithm similar to the one described in the AUTORG section. First, the input data $I(s)$ *vs.* s are converted into logarithmic scale. Then enumeration of all possible overlapping ranges is performed: for each range longer than a given minimum one data set is scaled against the other, the absence of systematic deviations is verified and discrepancy is computed as a root-mean-square deviation. Enumeration is performed in given steps to speed up the process. Each range is rated according to its length (number of points) and discrepancy, the range with the best rating providing the best overlap of the two data sets. In case of extrapolation to "zero" concentration the data from angles lower than the range with the best rating is extrapolated taking into account the given concentrations and the scaling factor for the selected range, assuming linear dependence of R_g on concentration (Fig. 3.4). Usually more than two concentrations are measured; in this case the highest concentration data is chosen as reference and extrapolations of each lower concentration data are performed against it; the resulting curves are averaged using the DATAVER tool (described in the next section). This is done by the automated data analysis pipeline as described below.

Note that the algorithm works also without prior normalization against concentration, in such case the scaling factor serves as a "relative concentration" value. This is particularly useful when the sample concentrations can not be easily determined e.g. due to lack of tryptophanes in a protein or in case of protein-nucleic acid complexes. Nevertheless the resulting extrapolated curve should be normalized in order to produce the correct $I(0)$ value.

In cases when concentration-dependant oligomerization or association-disassociation processes take place (e.g. Fig. 5.2, red circles) the use of ALMERGE obviously does not make sense although the algorithm will produce a composite curve. Therefore prior to merging or extrapolating to infinite dilution the linearity of the R_g *vs. concentration* plot should be checked.

ALMERGE was tested on numerous data sets (see *Application examples*) and the results were compared with those of manual merging done by the users. In all cases

the program yielded composite curves comparable or better than those obtained by the users. None of the users managed to correctly perform extrapolation to "zero" concentration, thus introduction of a tool like ALMERGE is important not only for the automated data analysis pipeline but also for manual/semi-automated data processing.

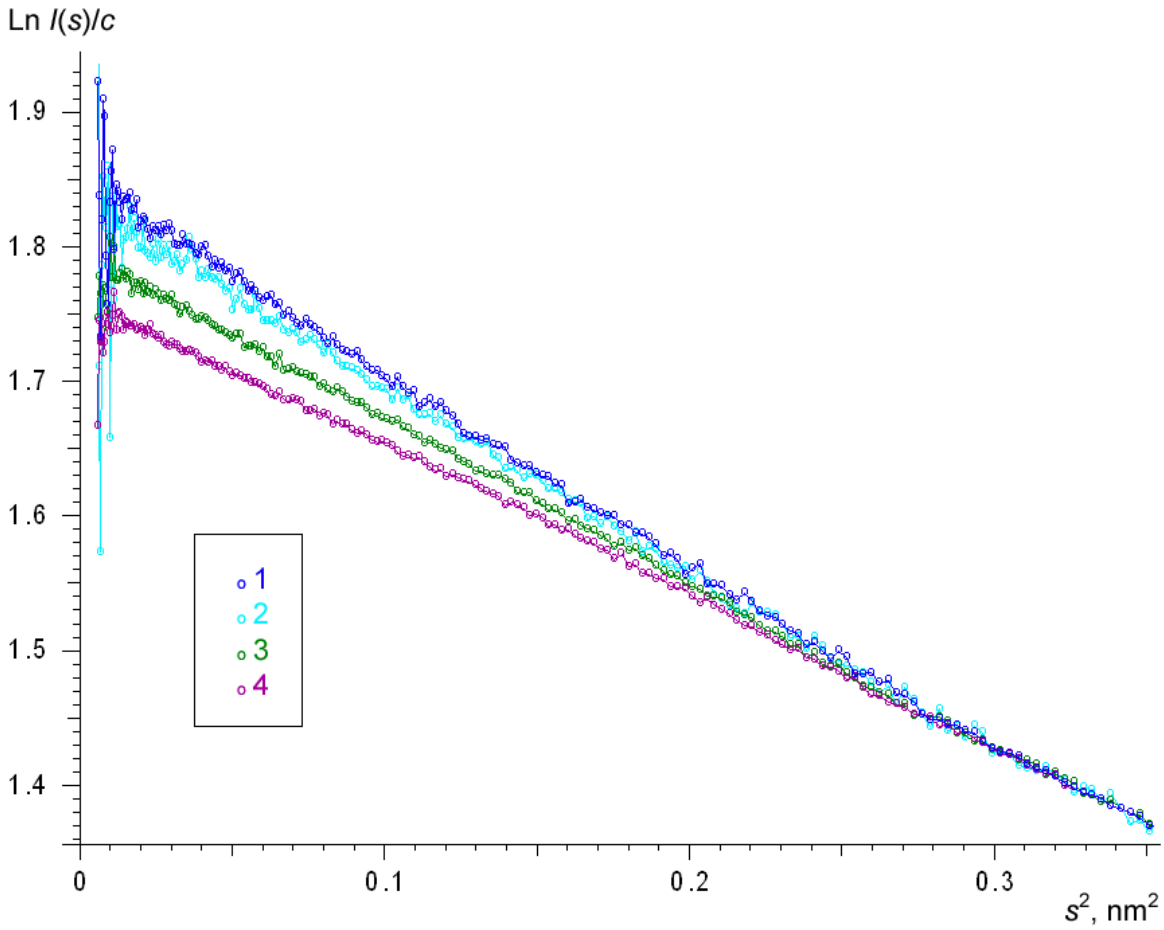


Fig. 3.5 Automated extrapolation to zero concentration of the scattering curve for the YtvA protein: (1) "infinite dilution" c = 0, (2)-(4) denote concentrations c = 2 mg/ml, 6 mg/ml, 10 mg/ml,

3.4 Modularized data processing tools

In the available SAXS data analysis packages single operations are directly linked to the master application providing limited or no scripting abilities, which makes it difficult to change sequence of steps in data processing schemes. To design a flexible automated data processing approach we introduce the concept of modularized tools:

- a modularized tool is essentially a primitive console application with a strictly defined interface;
- each data processing step can be performed using one tool or a set of tools; every tool can perform only one basic task;
- a tool can not call another tool; if certain input data is needed the caller should provide this data (presumably obtaining it from another tool in advance);
- to ensure reproducibility of obtained results each tool should provide its version number on request.

The range of possible application of tools is unlimited: tools can be reused by interactive GUI applications (e.g. PRIMUS), for manual data processing, in custom scripts etc. AUTORG, ALMERGE and the tools described below are used by the current implementation of the automated SAXS data analysis pipeline as explained in the next chapter.

3.4.1 DATCMP

Comparison of two data sets is a routine procedure during data analysis. Typically, two buffers are compared to inspect the stability of the background subtraction, or two different time frames are checked for radiation damage. This can be done with the primitive tool DATCMP:

```
$> datcmp [OPTIONS] <FILE> <FILE>
```

where the options are:

- -v; --version: print version information and exit.
- -h, --help: print a summary of arguments and options and exit.

Instead of a filename, one of the <FILE> arguments may be '-' to read data from stdin. The discrepancy between the two input files is calculated as root-mean-square deviation and printed to stdout. Note that in general, objective comparison of two scattering patterns to determine the presence or absence of systematic deviations

between them is a non-trivial task. However, the simple discrepancy criterion employed in DATCMP was proven to work well for the two aforementioned processing cases.

3.4.2 DATAVER

This tool is used to average two or more data sets into one:

```
$> dataver [OPTIONS] <FILES>
```

where the options are:

- -o, --output <FILE>: relative or absolute path to save the result; if not specified, the result is printed to stdout.
- -v; --version: print version information and exit.
- -h, --help: print a summary of arguments and options and exit.

Instead of a filename, one of the <FILE> arguments may be '-' to read data from stdin.

3.4.3 DATOP

DATOP is a tool for arithmetic operations on SAXS data. Primarily it is used for buffer subtraction and for scaling against concentration. It also can be used for the subtraction of a constant term. Usage:

```
$> datop <OPERATOR> <FILE> <FILE|X> [OPTIONS]
```

where the options are:

- -o, --output <FILE>: relative or absolute path to save the result; if not specified, the result is printed to stdout.
- -v; --version: print version information and exit.
- -h, --help: print a summary of arguments and options and exit.

Argument <OPERATOR> may be one of: ‘ADD’, ‘SUB’, ‘MUL’, ‘DIV’ (case-insensitive). Instead of a filename, one of the <FILE> arguments may be '-' to read data from stdin. The second <FILE> argument may also be a numeric constant X.

3.4.4 DATCROP

This tool is used to remove parts of the experimental data:

```
$> datcrop [OPTIONS] <FILE>
```

where the options are:

- -s, --skip <NUMBER>: how many first data points to crop; normally the first Guinier point minus one.
- -l, --last <NUMBER>: index of last point to be kept.
- --smin <NUMBER>: minimal *s* value to be kept.
- --smax <NUMBER>: maximal *s* value to be kept.
- -o, --output <FILE>: relative or absolute path to save the result; if not specified, the result is printed to stdout.
- -v; --version: print version information and exit.
- -h, --help: print a summary of arguments and options and exit.

To read data from stdin the <FILE> argument may be '-' instead of a filename.

3.4.5 DATGNOM

DATGNOM is a tool for automated estimation of the maximum intramolecular distance D_{max}, computation of the distance distribution function $p(r)$ and of the smooth regularized scattering curve (which is the Fourier transform of $p(r)$, see Equation 1.1 and Equation 1.2). This output is used by *ab initio* modeling programs; besides, the regularized curve is useful for the Porod volume estimation. Usage:

```
$> datgnom [OPTIONS] <FILE>
```

where the options are:

- -o, --output <FILE>: relative or absolute file path to save result in GNOM format.
- -r, --rg <NUMBER>: radius of gyration from the Guinier approximation.
- -s, --skip <NUMBER>: how many first data points to skip, normally the first Guinier point minus one.
- -v; --version: print version information and exit.
- -h, --help: print a summary of arguments and options and exit.

DATGNOM obtains the $p(r)$ function from an indirect Fourier transform of the scattering profile using a regularization procedure implemented in the program GNOM (Svergun 1992). In the original version of GNOM the maximum particle size D_{max} is a user-defined parameter and iterative calculations of $p(r)$ with different values of D_{max} are required to select its optimum value. This optimum D_{max} should provide a smooth real-space distance distribution function $p(r)$ such that $p(D_{max})$ and its first derivative $p_0(D_{max})$ are approaching zero, and the back-transformed intensity from the $p(r)$ (the "regularized curve") fits the experimental data. In DATGNOM, multiple GNOM runs are performed to find optimum D_{max} and $p(r)$ function. For globular particles D_{max} can be roughly estimated as three times the radius of gyration, e.g. for a hollow sphere $D_{max} = 2R_g$, for a solid sphere $D_{max} = 2.58R_g$, for an infinitely thin rod $D_{max} = 3.46R_g$. Thus the D_{max} values ranging from $2R_g$ to $3.5R_g$ are scanned with a step of $0.1R_g$, where R_g is provided by AUTORG. The calculated $p(r)$ functions for different D_{max} and corresponding fits to the experimental curves are compared using the perceptual criteria of GNOM, where the smoothness of $p(r)$, absence of systematic deviations in the fit and other quantities characterizing the result are merged into a total quality estimate. Moreover, the appropriately normalized value of $p_0(D_{max})$ is added to the estimate to ensure that the $p(r)$ function goes smoothly to zero. If no satisfactory solution is found the search continues for higher D_{max} values taking into account that the Rg value calculated from the slope of the regularized curve should be close to the entered R_g value. The highest possible D_{max} value is

limited to $D_{max} = 2\pi/s_{min}$. The best solution according to DATGNOM is selected and the $p(r)$ function together with the overall parameters is stored in the GNOM file format. The optimal D_{max} and the real-space R_g and $I(0)$ values are printed to stdout. Test computations with DATGNOM on various systems demonstrated that the program is able to reliably select the maximum size and calculate the $p(r)$ function, with results compatible with those of interactive analysis performed by an experienced user.

3.4.6 DATPOROD

The excluded volume of the particle can be computed using the Porod Equation 1.3. Application of this formula directly to the experimental data can be confronted with difficulties related to high noise level at higher angles; therefore the smooth regularized curve produced by DATGNOM is used:

```
$> datporod <GNOMFILE(S)> [OPTIONS]
```

where the options are:

- -v; --version: print version information and exit.
- -h, --help: print a summary of arguments and options and exit.

The Porod law assumes globular scatterers of uniform density (Porod 1951); thus the fact that the high angle part is discarded by DATGNOM is beneficial because it contains the contribution of internal particle structure which affects the accuracy of volume calculation (Glatter & Kratky 1982). An appropriate constant K_4 is subtracted from each data point to force the s^{-4} decay of the intensity. The truncation effect (integration up to a finite upper limit of s in the Porod equation) is taken into account as described in (Rolbin *et al.* 1973). The volume estimate is printed to stdout.

3.4.7 Reused tools

The above mentioned tools were either written from scratch or significantly modified to meet the pipeline requirements. In addition to them, there is a number of existing console applications, which can be used in the pipeline as is without adaptations.

The low resolution *ab initio* shape reconstruction program DAMMIF (Franke & Svergun 2009) represents the particle as a collection of several thousands of densely packed beads and employs simulated annealing to search for a compact interconnected model that fits the low-resolution portion of the data (usually to about 2 nm resolution) minimizing discrepancy:

$$\chi^2 = \frac{1}{N-1}\sum_j \left[\frac{I(s_j) - cI_{calc}(s_j)}{\sigma(s_j)}\right]^2,$$

where N is the number of experimental points, c is a scaling factor and $I_{calc}(s_j)$ and $\sigma(s_j)$ are the calculated intensity and the experimental error at the momentum transfer s_j, respectively. Recent developments allowed to significantly speed up the shape determination compared to other *ab initio* programs: obtaining a rough model that consists from a reduced number of beads takes only several minutes on a modern PC, a refined model is obtained in less than one hour. DAMMIF reads in the files in GNOM format, which can be automatically obtained from the SAXS data by means of AUTORG and DATGNOM. The output file in PDB format contains the model and its volume which can be used as an independent estimate of the molecular mass: for globular proteins, the bead model volume in nm^3 is about twice the molecular mass in kDa.

Due to the intrinsic ambiguity of the SAXS data interpretation, multiple DAMMIF runs may produce somewhat different models yielding nearly identical scattering patterns. These models can be superimposed and compared to obtain the most probable and an averaged model, which is done automatically in the program package DAMAVER (Volkov & Svergun 2003) which repeatedly uses the program SUPCOMB (Kozin & Svergun 2001) to align and compare pairs of models represented by beads. Such analysis allows one to assess the reliability of the *ab initio* models: multiple reconstructions from same input data should not demonstrate significant deviations from each other. This is expressed in terms of the normalized spatial discrepancy (NSD) output parameter.

An alternative higher resolution *ab initio* model can be constructed using the higher angles of scattering data by the program GASBOR (Svergun *et al.* 2001) representing

the protein as an ensemble of dummy residues forming a "chain-compatible" model. The spatial positions of these residues aim at approximately corresponding to those of the Cα atoms in the protein structure. The number of residues should be equal to that in the protein; if not known *a priori*, it is estimated from the calculated molecular mass. For large proteins, GASBOR is significantly slower than DAMMIF, therefore it should be started only if sufficient resources are available. Alternatively GASBOR can be started remotely over HTTP.

3.4.8 Online services

DARA is a database for rapid search of structural neighbors based upon their SAXS patterns (Sokolova *et al.* 2003). It is used to rank agreement between experimental data and scattering curves ($s < 1.5$ nm^{-1}) calculated from known high resolution protein structures which allows rapid search of structural neighbors at low resolution. A regularized curve (produced by DATGNOM) and the molecular mass are the required input parameters for DARA. A request to the database can be sent over HTTP, the result is obtained as an HTML page. A class that can send HTTP requests and wait for response in a separate thread (in parallel with other tasks performed by the pipeline) is reused from the cell filling control module described in chapter 2. This class can be used to also access other web-based services, e.g. DAMMIN or GASBOR which are available as an ATSAS-online application (Petoukhov *et al.* 2007).

3.5 Results

A set of tools for automated SAXS data analysis allows for rapid sample characterization and provides the overall parameters including radius of gyration, molecular mass, excluded volume and maximum particle dimension. The tools for data manipulation encompass all major data reduction steps ranging from basic buffer subtraction to infinite dilution extrapolation. The modularized approach allows developing flexible data processing scenarios that can include tools for automated 3D model building.

4 Automated data analysis pipeline

Despite the attempts of automation of some of the SAXS data processing steps (Petoukhov *et al.* 2007), (Hura *et al.* 2009) the entire process from the primary data reduction to 3D model building was not fully automated until recently. This is presumably due to the difficulties in obtaining reliable overall parameters from the experimental curves without user intervention. In the previous chapter we have introduced a number of tools for automated data analysis and manipulation; here we present an automated, robust, flexible SAXS data analysis pipeline based on these tools. The pipeline covers major processing and interpretation steps from primary data reduction to low-resolution three-dimensional model building.

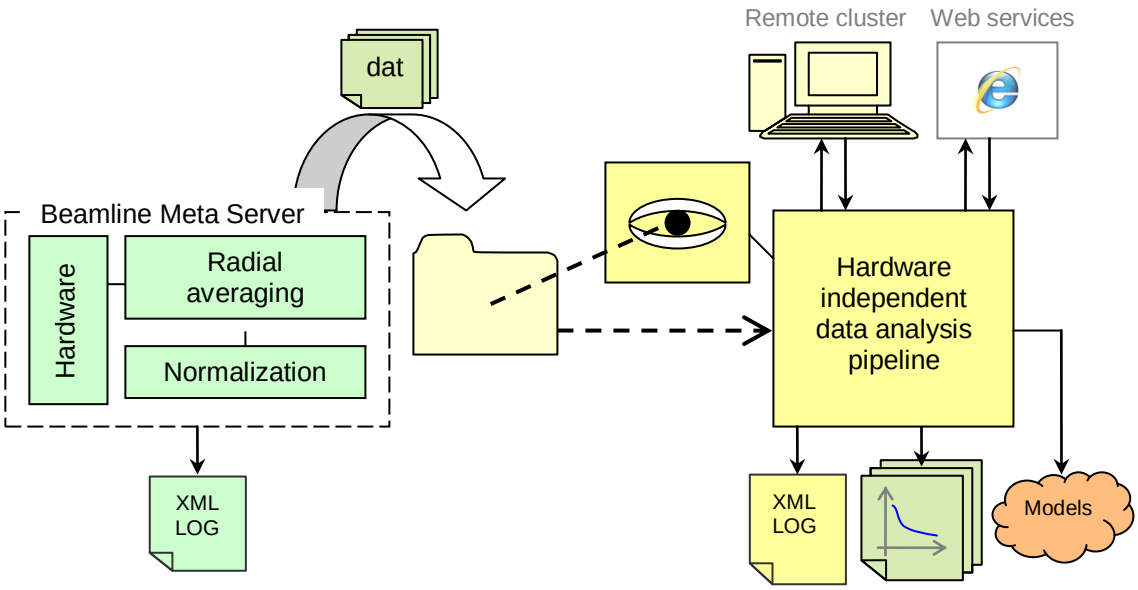


Fig. 4.1 Data analysis pipeline decoupled from the hardware control

4.1 Requirements

The primary data reduction is closely related to the instrument (hardware) design and setup. An ideal automated data analysis system should be hardware independent; therefore, the procedures like radial averaging and scaling against the primary beam intensity are beyond the scope of the present work. The data analysis pipeline is supposed to work in real-time to provide immediate feedback to the user during the measurements. Since remote beamline operation becomes a standard feature, the pipeline should allow remote control and easy network/web access to up-to-date results along with the experimental data. At the same time the pipeline should be sufficiently flexible for off-line batch reprocessing of the stored data if needed. Note that the batch processing does not require real-time delivery of intermediate results since all measured data is accessible at any step.

A possibility should also exist to switch off certain steps of the pipeline, which broadens the range of possible applications. Indeed, instead of performing all steps from buffer subtraction to *ab initio* shape determination one may need only to calculate the overall parameters and $p(r)$ functions from subtracted data without extrapolation to infinite dilution and 3D-modeling. On the other hand, new methods are being constantly developed, hence scalability, i.e. easiness of adding new steps is important.

More technical implementation requirements include parallelization and the ability to use remote nodes for distributed computing. From the programming point of view the pipeline should be cross-platform, i.e. compliable under major operating systems (Windows, Linux/UNIX variations, Mac OS). A comprehensive data analysis system is being implemented by multiple developers who may use more than one programming language (e.g. reusing existing code). A separation of responsibilities between individual operations is important both for debugging and for tracking down possible artifacts. Stability against software bugs, user mistakes and hardware failures should ensure that no data is lost and no incorrect results are produced.

4.2 Possible implementation approaches

The most obvious approach would be to create a single application that performs all data processing operations, but such an application is extremely hard to maintain. A somewhat more structured approach involves using shared binary libraries, e.g. dynamic-link libraries in Windows. Such libraries contain reusable routines which are linked in the runtime of the calling application decreasing the binary sizes and permitting usage of more than one programming language. Apart from well-known DLL problems related to possible backward incompatibility of different library versions (so-called "DLL hell") repeated usage of same library on same data is an unwanted feature; e.g. each time the R_g value is needed the AUTORG DLL is called which in a general case can lead to inconsistent state of processed data.

Instead of using binary libraries one could also subdivide data processing steps into console programs and execute them sequentially using a shell script (*.bat files on Windows). This more flexible approach has a drawback related to the real-time requirement: execution time of different steps varies significantly, e.g. buffer subtraction takes milliseconds whereas *ab initio* shape determination may take minutes or even hours preventing the script from doing anything else. Besides, scripts are not portable. A real-time-oriented approach should allow for asynchronous usage of data processing modules. Each console program could be wrapped into a generic class that is running the program in a separate thread, saving the result to a common XML log file or a relational database from where other classes can read it. Tests have shown that managing asynchronous reading/writing using a mutex (mutual exclusion) object results in a considerable synchronization overhead.

To meet the foregoing requirements and to avoid the above mentioned drawbacks we have developed a pipeline for automated hardware-independent SAXS data analysis that combines the advantages of a modularized approach with the ability to run console applications in separate threads controlling the running state and output. Thus the pipeline is implemented as a separate "master" application that manages serial or concurrent execution of command-line tools without performing any data analysis apart from decision-making.

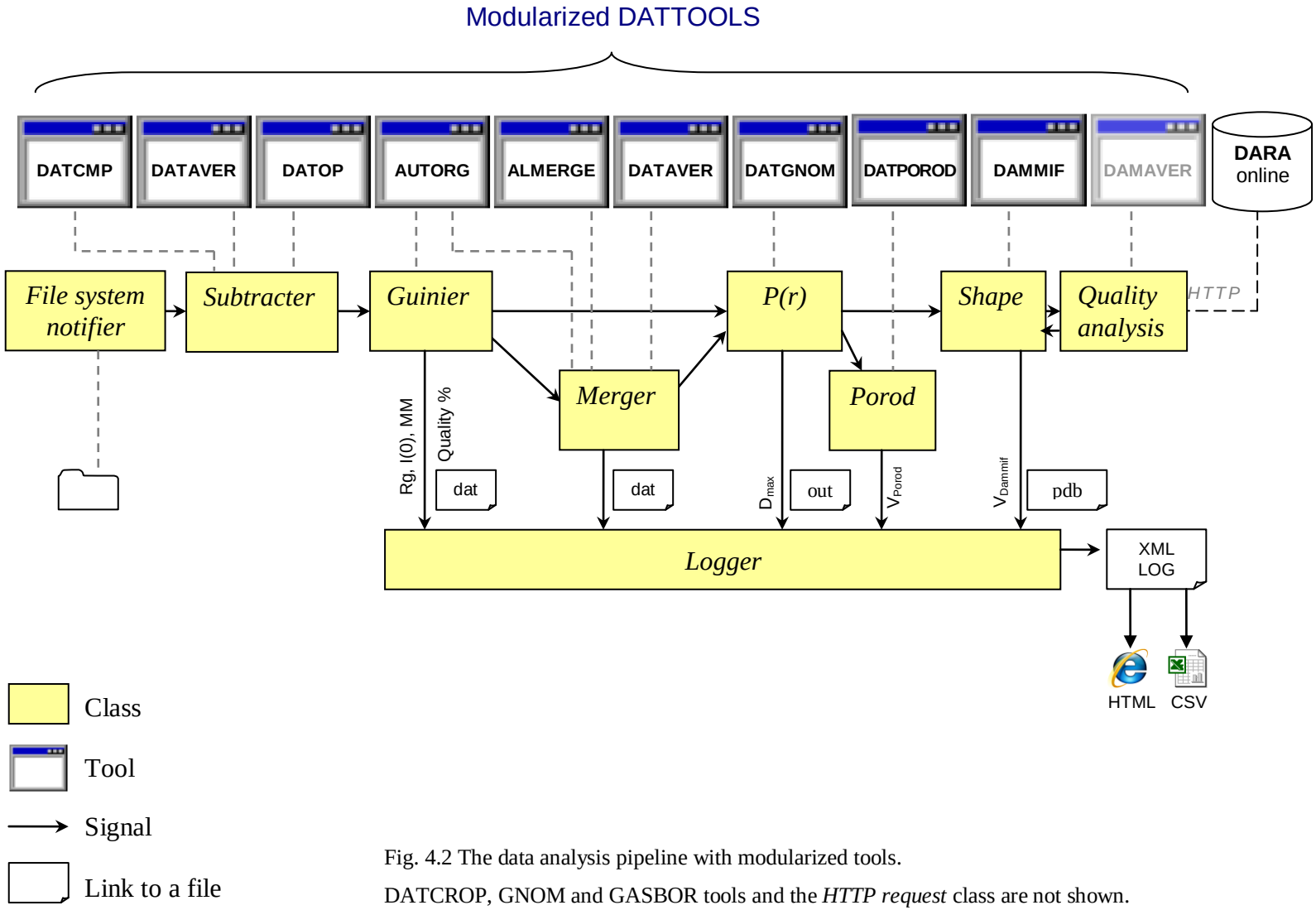


Fig. 4.2 The data analysis pipeline with modularized tools.

DATCROP, GNOM and GASBOR tools and the *HTTP request* class are not shown.

4.3 Pipeline implementation

Each data processing step is implemented as a separate C++ class that employs one or more of the modularized tools to perform operations on the data (Fig. 4.2). In such a way the classes are only decision-making blocks that do not perform any actual data processing. Classes communicate with each other using the *signals and slots* mechanism provided by the Qt framework (http://qtsoftware.com). A signal is emitted when a particular event occurs: a new file is available for processing, a tool has finished calculations etc. A slot is a function that is called in response to a particular signal. Such mechanism ensures that if a signal is connected to a slot, the slot will be called with the signal's parameters at the right time. One signal can be received by several slots; a class which emits a signal neither knows nor cares which slots receive the signal. This way of connecting different classes enables one to modify the behavior of the pipeline to meet different requirements, e.g. by excluding certain steps if needed. Introducing new classes that will perform any new data processing operations consequently or in parallel with the existing setup is made particularly easy.

Radially averaged normalized one-dimensional data sets where intensity is a function of the scattering vector are the expected input to the pipeline. The distinction between the solute (sample) and the solvent (buffer) data is made by the concentration value (buffers have zero concentration). Each sample is expected to have its code that identifies the specimen along with its buffer and preparation particularities, i.e. separate samples should have different codes, a concentration series of same sample has the same code. The *libsaxsdocument* library is attempting to read the intensity values along with their experimental error values, sample code, concentration, description and several other parameters from the headers of the reduced files. If the sample code can not be read, an attempt to extract it from file name is made, whereas failure to read the concentration value will make further data processing impossible: the buffers are distinguished from the samples by having zero concentration. It is assumed that the measurement of the sample is surrounded by two buffer

measurements (not necessarily consecutive) and the sample can be exposed more than once in order to monitor possible radiation damage.

The *File system notifier* class runs in a separate thread (in parallel to other classes), constantly checking the input folder for new reduced files. It has three settings: path to a local or remote directory that should be monitored for new files; file name mask; the time how often to check the folder contents. Theoretically this should be doable through the respective operating system service but tests show that such a solution is unreliable. Besides, the present implementation ensures better portability. The *File system notifier* class emits a signal every time a new file is detected.

If the received file is a sample, the *Subtracter* class subtracts the latest available buffer using DATOP to obtain a temporary result. As soon as the next buffer becomes available, the buffers measured before and after the sample are compared. DATCMP is used to characterize the stability of the buffers. Similarly, DATCMP is used to compare successive sample measurements with matching code and concentration values to monitor for possible radiation damage, if relevant. Depending on the settings, successive measurements of same sample can also be treated as independent measurements or only the data from the former measurement will be passed on to the next step. If the sample data differ beyond a certain threshold, the radiation damage flag is written to the sample properties. If the compared files are statistically indistinguishable, appropriate averaging operations are done with DATAVER and the averaged background is subtracted and scaled against concentration by DATOP. If not, then the individual buffers are subtracted along with the averaged buffer and sent to the *Guinier* class where the quality of the three subtractions is checked using AUTORG. The subtraction with the best quality is passed to the next steps along with the respective AUTORG output and molecular mass calculated using Equation 3.2 where M_{ref} is taken from the settings and $I_{ref}(0)$ is obtained from the reference sample measurement.

The signal from *Guinier* is received by three classes: *Logger, P(r)* and *Merger*.

The *Logger* class receives most of the signals from other objects. It collects new information about the data being processed and makes sure it is written to the XML log which is described in detail below. Since the log file might be stored on a remote

hard disk drive and accessed via network file system (NFS), *Logger* makes certain that the file is updated not too frequently; also it is taken into account that NFS is based on the UDP protocol which does not implement delivery confirmation. When the pipeline is started, the constructer of the class attempts to read the log file and to restore the last known consistent state.

The *P(r)* class first checks the quality of the data that was reported by AUTORG. If it indicates strong presence of aggregation, no further processing is performed. However, some samples show nonlinearity over only a small region of the Guinier plot. Data cropped at the lowest angles where aggregation is most apparent may be processed by DATGNOM. When the calculation is over the R_g value calculated from the $p(r)$ function (essentially using the information from the higher angles of the scattering curve) is compared to the R_g value of the Guinier approximation. If the difference is more than expected (5% by default) the DATGNOM output is ignored, otherwise a signal is emitted. The *Porod* class receives the signal from *P(r)* and computes the hydrated volume using DATPOROD.

The *Merger* class also checks the quality of the data and, if it is higher than a specified value, compares the sample code with previously measured samples. If more than two samples of satisfactory quality are available, the class attempts extrapolation to infinite dilution using ALMERGE. Prior to this the R_g-*concentration* dependence is checked: if R_g changes insignificantly with concentration or the coefficient of determination R^2 (Steel & Torrie 1960) of the linear approximation is too low, the extrapolation step is skipped. If no concentration effects are observed (R_g does not change), the highest concentration data is marked as final and passed further down the pipeline. The soundness of the extrapolated data is verified by comparing the R_g value derived from AUTORG to the one derived from the linear approximation. If the result is evaluated as unsatisfactory, the composite curve is deleted, otherwise the lowest angles before the first Guinier data point are cropped using DATCROP and a corresponding signal is emitted. If only two samples of satisfactory quality and different concentrations are available, they are merged by ALMERGE without extrapolating to infinite dilution. Every composite curve is processed by the *P(r)* class as described above.

The *Shape* class is managing the *ab initio* modeling programs, primarily DAMMIF. A single DAMMIF run in "fast" mode produces a rough shape along with a volume estimate in a few minutes; however it is important to avoid blocking the pipeline while calculations are in progress. Therefore all time-consuming tools are run in parallel in separate threads; when a calculation is over a signal is emitted. A single DAMMIF run is performed, according to the settings, either for every curve processed by the *P(r)* class or only for composite curves previously produced by *Merger*.

One of the challenges of automated data processing lies in determining an appropriate scoring function that successfully distinguishes correct results. Apart from data quality estimation implemented in AUTORG and check-ups in the *Merger* and *P(r)* classes, values obtained from different tools can be used to independently confirm the results. For that the *Quality analysis* class compares the molecular mass estimated from volume and $I(0)$ of "infinite dilution" data. Our experience shows that for globular proteins Porod volume in nm^3 is about 1.6 times the molecular mass in kDa; volume of a DAMMIF model in nm^3 is about twice the molecular mass in kDa. If variation of the three derived molecular mass values is less than 20%, then further, more computationally expensive data analysis steps are performed.

A good consistency check for the *ab initio* models is provided by DAMAVER. To assess the uniqueness of the low resolution shape DAMMIF is executed by the *Shape* class several times (normally six to twelve, depending on available computational resources). If the pipeline is running on a multiple-processor system, it is possible to perform all DAMMIF runs simultaneously; nonetheless in the real-time mode it should be avoided for two reasons: a) if all CPUs are already busy calculating models for one sample, starting the runs for another sample will significantly slow down the performance; b) if another concentration of the same sample was measured during the current DAMMIF run, the next run can be performed for the updated composite curve. Of course employing a remote cluster to decouple the real-time pipeline from computationally expensive tasks is preferable. If the NSD value computed by DAMAVER is less than 1 the model ranked as most probable is saved as the final one.

Another shape determination approach developed specifically for proteins is implemented in the GASBOR program. Curiously enough when a nucleic acid is measured, it does not pass the molecular mass comparison check: since nucleic acids have approximately two times higher contrast compared to proteins, the molecular mass obtained from the $I(0)$ value does not agree with the other two estimates. When the molecular mass values agree the number of residues is estimated from the molecular mass assuming that one residue weighs 110 Da. One should note, however, that the number of residues is typically known *a priori* from the amino acid sequence of the protein, and GASBOR is usually run as a separate application. GASBOR can not handle more than 6000 residues per asymmetric unit; however it is recommended to further limit the maximum size of the protein because the execution time grows quadratically with the number of residues. To take higher angles into account the input file for GASBOR is generated by running GNOM on the full-length data with the D_{max} and *alpha* parameters previously estimated by DATGNOM. Currently GASBOR is not started by default because it is too resource-intensive but we believe that in the nearest future either it will be optimized or the available computational power will increase.

The *HTTP request* class also benefits from running in parallel to the main process. It reuses the mechanism employed in the cell filling control module (described in chapter 2) to send requests to the DARA database. The request contains the molecular mass value and a file in GNOM format; the response is an HTML page that can be parsed relatively easily. Using this class one can utilize other tools available via the HTTP protocol, e.g. GASBOR, DAMMIN, DAMMIF/DAMAVER.

4.4 Storage of the results

It is vital for automated procedures dealing with large amounts of data that both the obtained information and history of the data analysis (including the scattering curves and the obtained results) are easily retrievable in human- and machine-friendly forms. To achieve convenient hierarchical data storage and report generation, XML-based file format was developed for summary/log files. XML stands for Extensible Markup Language; it is a general-purpose specification for creating custom markup languages. XML was designed as a simplified subset of the Standard Generalized Markup

Language (SGML, ISO 8879) and became a recommendation of the World Wide Web Consortium (W3C) in 1998 (http://www.w3c.org/xml). Initially defined for large-scale electronic publishing, XML dialects nowadays are also employed in a wide variety of data exchange applications. The data in an XML dialect are stored in plain text files of hierarchical human readable structure (Fig. 4.3, left). A multitude of matured parsers and translation software tools for XML data files are available, both proprietary and open source. XML is one of the technologies providing solutions for sharing information across different computing platforms and thus presents a practical approach to data categorization and communication. The ability to create a custom tagging structure gives the language the possibility to categorize and structure data for both ease of retrieval and ease of display.

After each data processing step the *Logger* class receives the corresponding signal and a summary of results such as R_g, molecular mass, quality estimation, D_{max}, volume etc. is immediately written to the XML-log granting consistency of the log at any given time point. If the work of the pipeline was interrupted, e.g. because of a hardware failure, on restart the pipeline looks for the log file and attempts to read it in order to recover the last known consistent state. If the log is not found or corrupt, the pipeline will reprocess all available data and create a new log file.

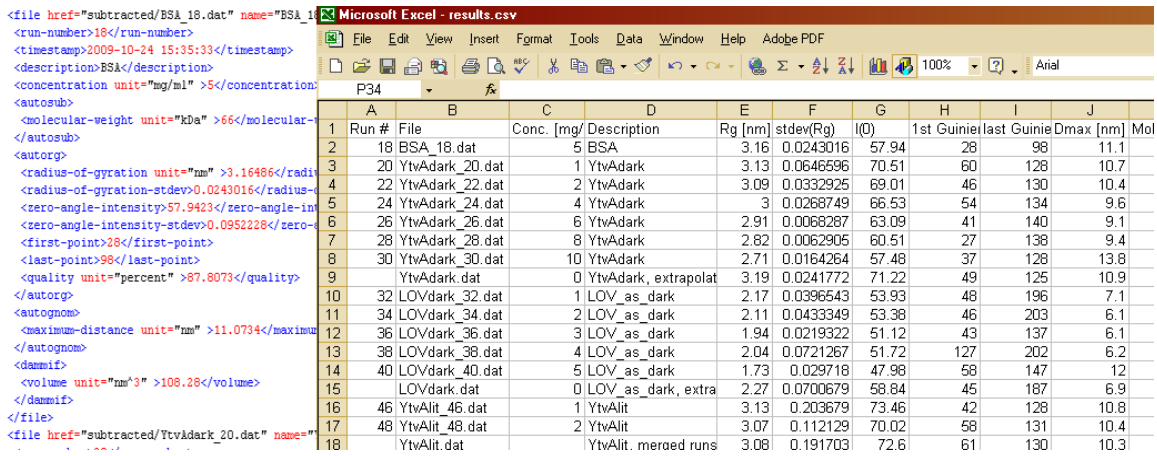
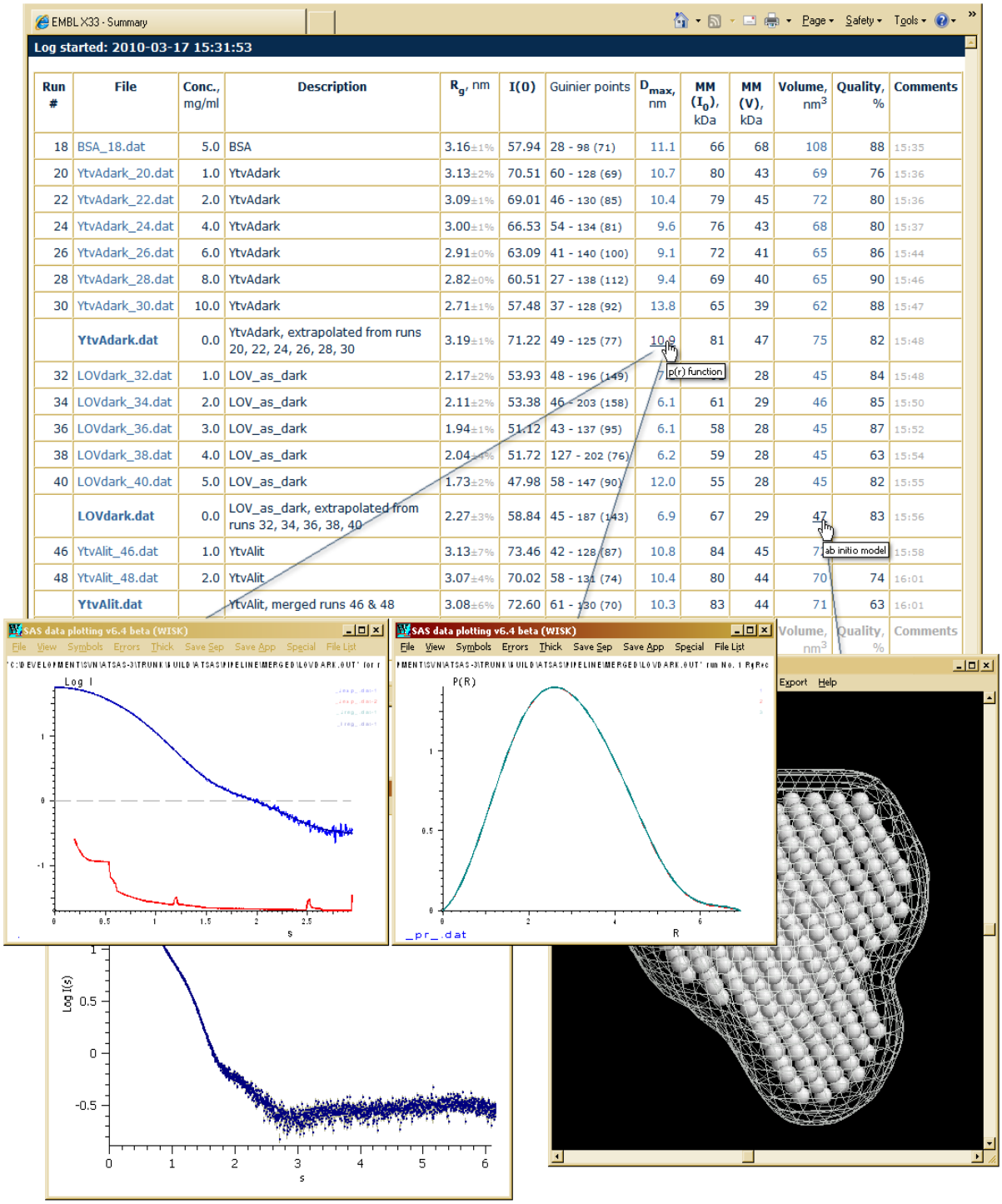


Fig. 4.3 The XML source code opened in a text editor (left) and the same code XSL processed into CSV format and opened with Excel (right).

The format allows for easy data processing using standard XSL transformations – web-friendly HTML representation and export to CSV (e.g. for use in Excel, see Fig. 4.3, right); conversion to other formats is possible. Thus the summary and logbook files can be opened by a conventional HTML browser e.g. Internet Explorer or Firefox. The XML data is converted into a human-readable form on the fly and is presented as an HTML table with links to the data files and models (see Fig. 4.4). The Web-accessible format permits the use of the log for remote control of the measurements facilitating access to the reduced data files and estimated overall parameters. To open the subtracted *.dat file one needs to click to the file name; a click on the D_{max} value opens the *.out file that contains the p(r) function; the molecular mass value is linked to DARA results page; a click on the volume opens the low resolution DAMMIF model.

The described format is used at the X33 beamline since October'2008. An automatic logbook of monitor values such as cell temperature and current beam characteristics is also written in a similar XML format to a separate file.



Run #	File	Conc., mg/ml	Description	R_g, nm	I(0)	Guinier points	D_{max}, nm	MM (I_0), kDa	MM (V), kDa	Volume, nm^3	Quality, %	Comments
18	BSA_18.dat	5.0	BSA	3.16±1%	57.94	28 - 98 (71)	11.1	66	68	108	88	15:35
20	YtvAdark_20.dat	1.0	YtvAdark	3.13±2%	70.51	60 - 128 (69)	10.7	80	43	69	76	15:36
22	YtvAdark_22.dat	2.0	YtvAdark	3.09±1%	69.01	46 - 130 (85)	10.4	79	45	72	80	15:36
24	YtvAdark_24.dat	4.0	YtvAdark	3.00±1%	66.53	54 - 134 (81)	9.6	76	43	68	80	15:37
26	YtvAdark_26.dat	6.0	YtvAdark	2.91±0%	63.09	41 - 140 (100)	9.1	72	41	65	86	15:44
28	YtvAdark_28.dat	8.0	YtvAdark	2.82±0%	60.51	27 - 138 (112)	9.4	69	40	65	90	15:46
30	YtvAdark_30.dat	10.0	YtvAdark	2.71±1%	57.48	37 - 128 (92)	13.8	65	39	62	88	15:47
	YtvAdark.dat	0.0	YtvAdark, extrapolated from runs 20, 22, 24, 26, 28, 30	3.19±1%	71.22	49 - 125 (77)	10.0	81	47	75	82	15:48
32	LOVdark_32.dat	1.0	LOV_as_dark	2.17±2%	53.93	48 - 196 (149)	7		28	45	84	15:48
34	LOVdark_34.dat	2.0	LOV_as_dark	2.11±2%	53.38	46 - 203 (158)	6.1	61	29	46	85	15:50
36	LOVdark_36.dat	3.0	LOV_as_dark	1.94±1%	51.12	43 - 137 (95)	6.1	58	28	45	87	15:52
38	LOVdark_38.dat	4.0	LOV_as_dark	2.04±4%	51.72	127 - 202 (76)	6.2	59	28	45	63	15:54
40	LOVdark_40.dat	5.0	LOV_as_dark	1.73±2%	47.98	58 - 147 (90)	12.0	55	28	45	82	15:55
	LOVdark.dat	0.0	LOV_as_dark, extrapolated from runs 32, 34, 36, 38, 40	2.27±3%	58.84	45 - 187 (143)	6.9	67	29	47	83	15:56
46	YtvAlit_46.dat	1.0	YtvAlit	3.13±7%	73.46	42 - 128 (87)	10.8	84	45	7		15:58
48	YtvAlit_48.dat	2.0	YtvAlit	3.07±4%	70.02	58 - 131 (74)	10.4	80	44	70	74	16:01
	YtvAlit.dat		YtvAlit, merged runs 46 & 48	3.08±6%	72.60	61 - 130 (70)	10.3	83	44	71	63	16:01
										Volume, nm^3	Quality, %	Comments

Fig. 4.4 The XML log is automatically converted to HTML when opened in a web-browser. The entries in columns *File*, D_{max} and *Volume* are clickable.

4.5 Validation on simulated data

To test the pipeline on simulated data we took yeast bleomycin hydrolase (PDB ID 3GCB, (Zheng *et al.* 1998)), a hexameric protein used in the Protein interfaces, surfaces and assemblies service PISA as an example (Krissinel & Henrick 2007). From high resolutions structures of a monomer, a hexamer and two dimmers suggested by PISA (a compact one and an extended one) scattering curves were calculated using CRYSOL (Svergun *et al.* 1995) in the range of momentum transfer $0.13 < s < 3.1$ nm^{-1}; noise was added to simulate experimental errors (Fig. 4.5). A scattering curve from an aggregated sample was simulated as a mixture of 90% monomers and 10% 54-mers (Fig. 4.5, inset). The overall results are summarized in Table 4.1. The accuracy of the radius of gyration estimated by AUTORG and DATGNOM is about 3%. Although it was possible to estimate the R_g value for the "aggregated" curve, the quality check prevented the pipeline from further data processing. The molecular mass tends to be slightly underestimated but leaves no doubt about the oligomeric state. The D_{max} values of the larger constructs are closer to the ones of the models, although the computed $p(r)$ functions sometimes display ripples (Fig. 4.6). The reconstructed *ab initio* models are strikingly close to the initial shapes (Fig. 4.7).

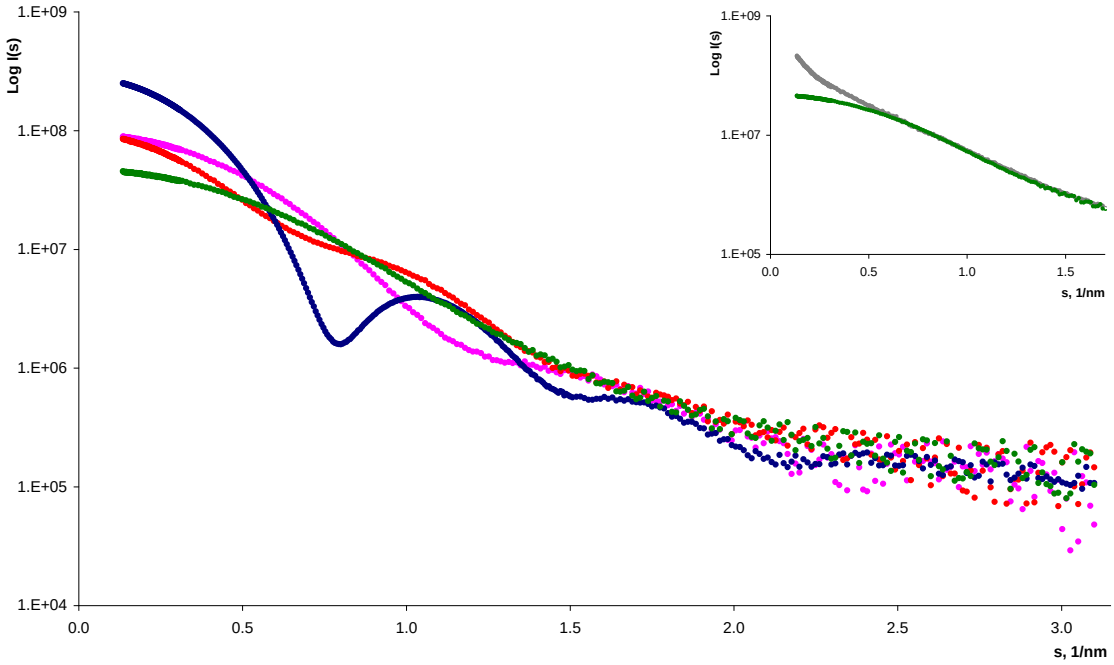


Fig. 4.5 Data simulated from the yeast bleomycin hydrolase structure: a monomer (green), a globular dimer (magenta), an extended dimer (red), a hexamer (blue) and an aggregated sample (inset, gray).

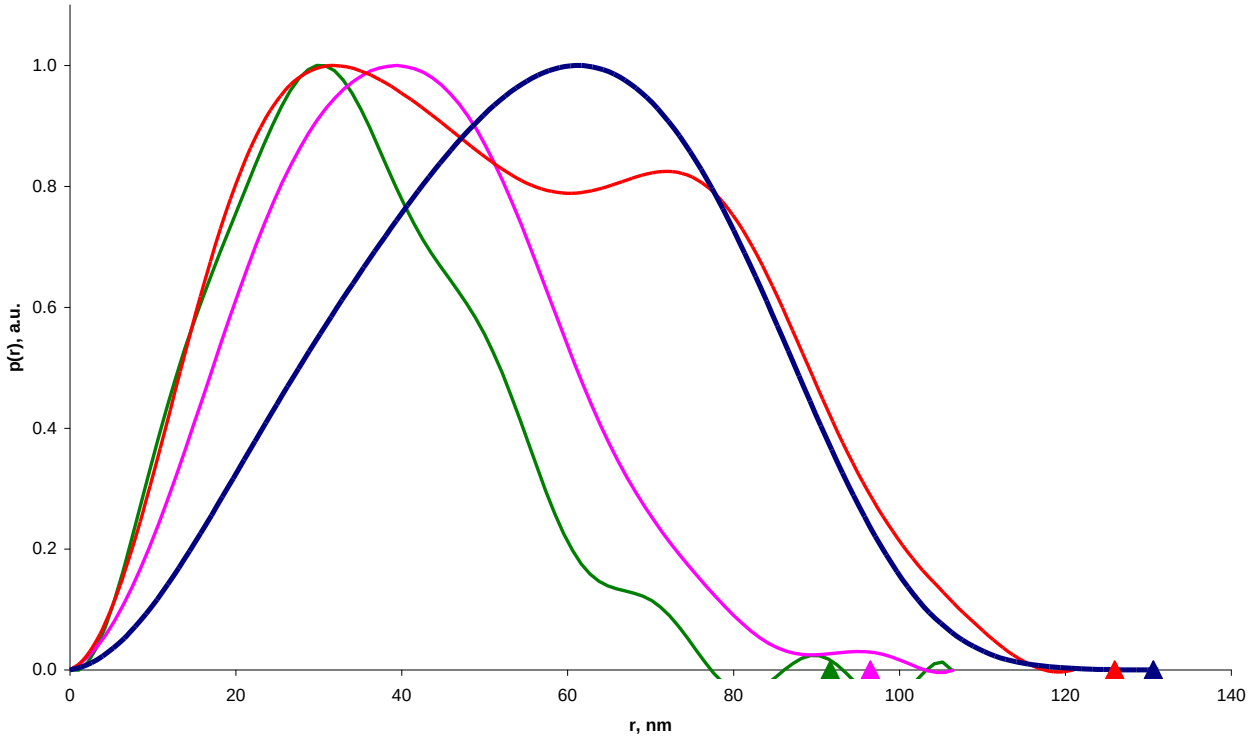


Fig. 4.6 p(r) functions obtained by the automated pipeline from data simulated from a monomer (green), a globular dimer (magenta), an extended dimer (red) and a hexamer (blue). The triangles on the abscissa axis mark the known theoretical D_{max}.

	Theoretical			AUTORG		DATGNOM		DATPOROD		DAMMIF	
	R_g, nm	D_{max}, nm	MM, kDa	R_g, nm	MM, kDa	R_g, nm	D_{max}, nm	V_p, nm^3	MM(V_p), kDa	V_{dam}, nm^3	MM(V_{dam}), kDa
3GCB Monomer	2.61	9.2	50	2.6±0.1	46±5	2.6	11	66	42	89	44
Globular dimer	3.07	9.6	101	3.1±0.1	92±9	3.1	11	143	90	172	86
Extended dimer	4.09	12.6	101	4.1±0.1	92±9	4.1	12	140	87	186	93
Hexamer	4.37	13.1	302	4.4±0.0	275±28	4.3	13	426	266	606	303
Aggregates, 50-2721 kDa				3.7± ∞	Further evaluation was not performed due to aggregation						

Table 4.1 Validation on simulated data: overall parameters

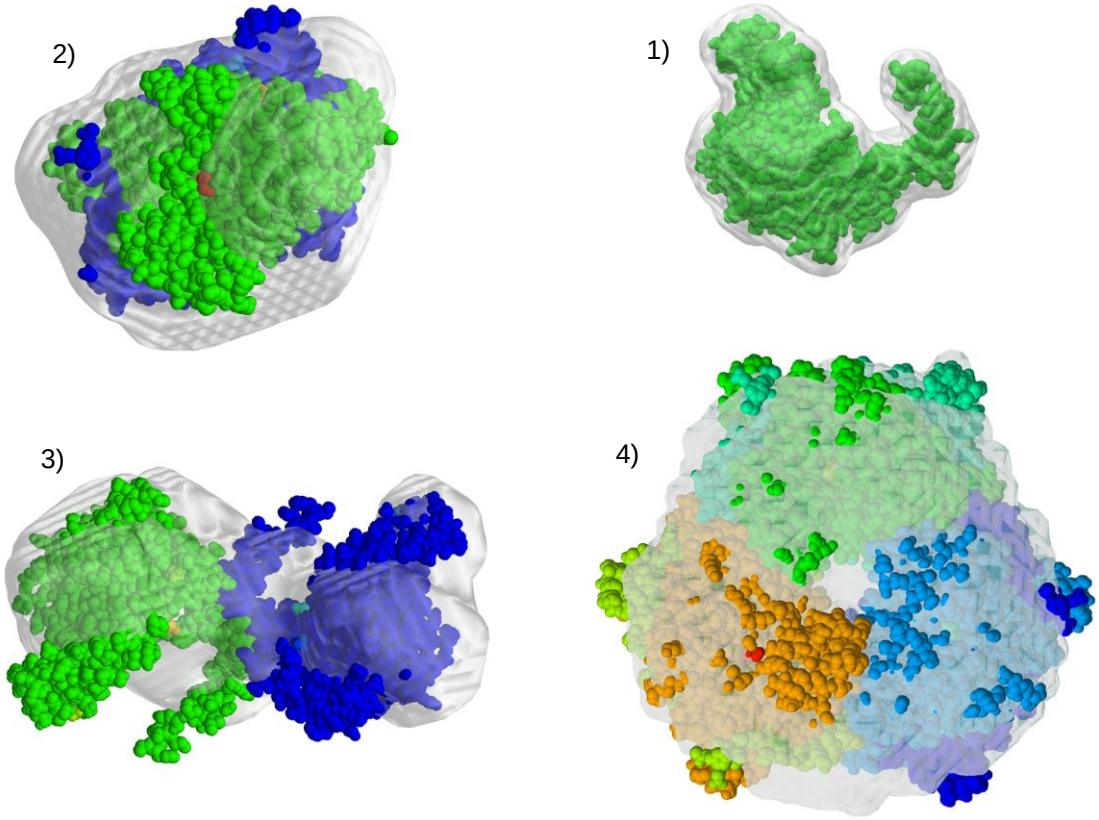


Fig. 4.7 Validation on simulated data: The automatically reconstructed *ab initio* models (gray) are strikingly close to the initial shapes (color). 1) monomer, 2) globular dimer, 3) extended dimer, 4) hexamer.

4.6 Results

The pipeline covers major processing and interpretation steps from primary data reduction to low-resolution three-dimensional model building and is capable of working with data from monodisperse or moderately polydisperse solutions of macromolecules. Quality control of input data and means of intermediate results validation offer for the first time the opportunity to get an expert feedback in a form of a structured summary of overall specimen parameters immediately after the measurement facilitating unattended and remote-controlled measurements. An object-oriented implementation approach combined with the modularized tools concept results in an adjustable, scalable, robust, portable solution.

5 Practical applications

The automated data processing pipeline was used to evaluate data from numerous user projects. In the following chapter three projects that deal with samples of various nature are described.

The synchrotron radiation X-ray scattering data from solutions of each sample were collected on the X33 camera of the EMBL on the storage ring DORIS III (DESY, Hamburg, Germany). Using a MAR345 image plate or PILATUS 1M photon counting detector at a sample-detector distance of 2.7 m and a wavelength of $\lambda = 0.15$ nm, the range of momentum transfer $0.1 < s < 5$ nm^{-1} was covered. For each construct, several solute concentrations in the range of 1 to 5-10 mg/ml were measured. To monitor for the radiation damage, two or four successive exposures of sample solutions were compared and no significant changes were observed. The data were normalized to the intensity of the transmitted beam and radially averaged. The automated pipeline subtracted the scattering of the buffer and scaled the difference curves for protein concentration; the low angle data measured at lower sample concentrations were extrapolated to infinite dilution and merged with the higher concentration data to yield the final composite scattering curves. Overall parameters such as radius of gyration, molecular mass (derived from $I(0)$), D_{max} and Porod volume were evaluated along with data quality estimation. The distance distribution function was computed, followed by low resolution shape determination using the *ab initio* program DAMMIF. To obtain an independent estimate of the molecular mass, the latter was also calculated from the Porod volume and from the volume of the DAMMIF models. These data processing steps were performed by the automated pipeline without user intervention.

5.1 Protein-RNA complexes: chemokines inhibited by spiegelmers

Chemokines are small proteins stabilized by disulfide bridges, containing 60-70 amino acids that form a subfamily of cytokines with a high level of chemotactic activity as a response to inflammatory signals. Chronic inflammatory diseases such as diabetic nephropathy or lupus nephritis are associated with increased expression levels of chemokines (Ruster & Wolf 2008), (Tucci *et al.* 2009). Furthermore, chemokines play an important role in hematopoietic disorders (Broxmeyer 2008).

In order to reduce inflammatory reactions or to influence hematopoiesis of the patients NOXXON Pharma AG developed therapeutic RNA structures, so called spiegelmers that bind and inactivate specific chemokines. Spiegelmers represent a new class of RNA structures (L-aptamers) synthesized by mirrorimage nucleotides using a patented technology developed by NOXXON Pharma AG. Spiegelmers bind with high specificity and affinity to target molecules, e.g. chemokines, causing inactivation of the target. These therapeutic L-aptamers are characterized by high stability in human plasma without inducing an immune response in the body, which allows their application as a new class of pharmaceuticals (Maasch *et al.* 2008), (Sayyed *et al.* 2009). A better understanding of the mechanism of chemokine inactivation by spiegelmers is required to optimize the treatment of the associated diseases. Here we describe the structural interactions of the inhibitors NOX-A12 and NOX-E36 with the chemokines SDF-1 and MCP-1, respectively, investigated by SAXS.

First we analysed the structural interactions between the chemokine SDF-1 (stroma cell derived factor-1, CXCL12) and its corresponding spiegelmer NOX-A12. NOX-A12 was developed as an inducer of hematopoietic stem cell mobilization from bone marrow into peripheral blood. It is a picomolar inhibitor of SDF-1, a key regulatory element in the homing and retention of hematopoietic stem cells in the bone marrow. NOX-A12 is currently investigated in a phase I clinical trial which will be completed in the second quarter of 2010. NOXXON plans to start a phase II clinical trial with haematological cancer patients in the second half of 2010.

Moreover, we investigated the structural interactions between the chemokine MCP-1 (monocyte chemoattractant protein-1, CCL2) and its corresponding spiegelmer NOX-E36. MCP-1, a member of the CC subfamily of chemokines, acts as a chemoattractant for monocytes and is responsible for their migration to inflamed tissues. NOX-E36 is a picomolar inhibitor of MCP-1, implicated in a variety of inflammatory diseases including diabetic nephropathy, lupus nephritis, and rheumatoid arthritis. Studies in a lupus model indicate significant improvement in survival rates, kidney morphology, and other autoimmune-related symptoms in animals treated with NOX-E36. A phase I study is currently ongoing.

For the automated SAXS investigation of the structural interactions of the NOX aptamers with the respective chemokines we initially measured the scattering profiles of each complex component alone (NOX aptamers (1), SDF-1 and MCP-1 (2)) followed by measurement of the complex (3). The questions addressed were: is it possible to model the 3D structure of the NOX aptamers, which was so far unknown? Does the solution structure of the chemokines compare to the known crystal structures and NMR structures already deposited in the PDB database? What is the oligomeric state of the components alone and in complex? What is the ratio RNA:protein ratio present in the complex (1:1 or different)?

5.1.1 NOX-aptamers NOX-A12 and NOX-E36

The obtained molecular mass values both for NOX-A12 and NOX-E36 pointed to monomers in solution (see Table 5.1). Since the 3D structure of the NOX aptamers is unknown we used the RNA structure prediction server MC-Fold|MC-Sym (Parisien & Major 2008) to calculate 2000 possible models for each aptamer. The theoretical scattering from each model was calculated with CRYSOL (Svergun *et al.* 1995). Given the atomic coordinates, this program minimizes discrepancy in the fit to the experimental intensity by adjusting the excluded volume of the particle and the contrast of the hydration layer. The models with the lowest discrepancy yielded χ values below 1 and were chosen as the final 3D structures. Thus, by combining the information obtained from the SAXS profile and from the structure prediction server we obtained the most likely 3D model of each aptamer.

-

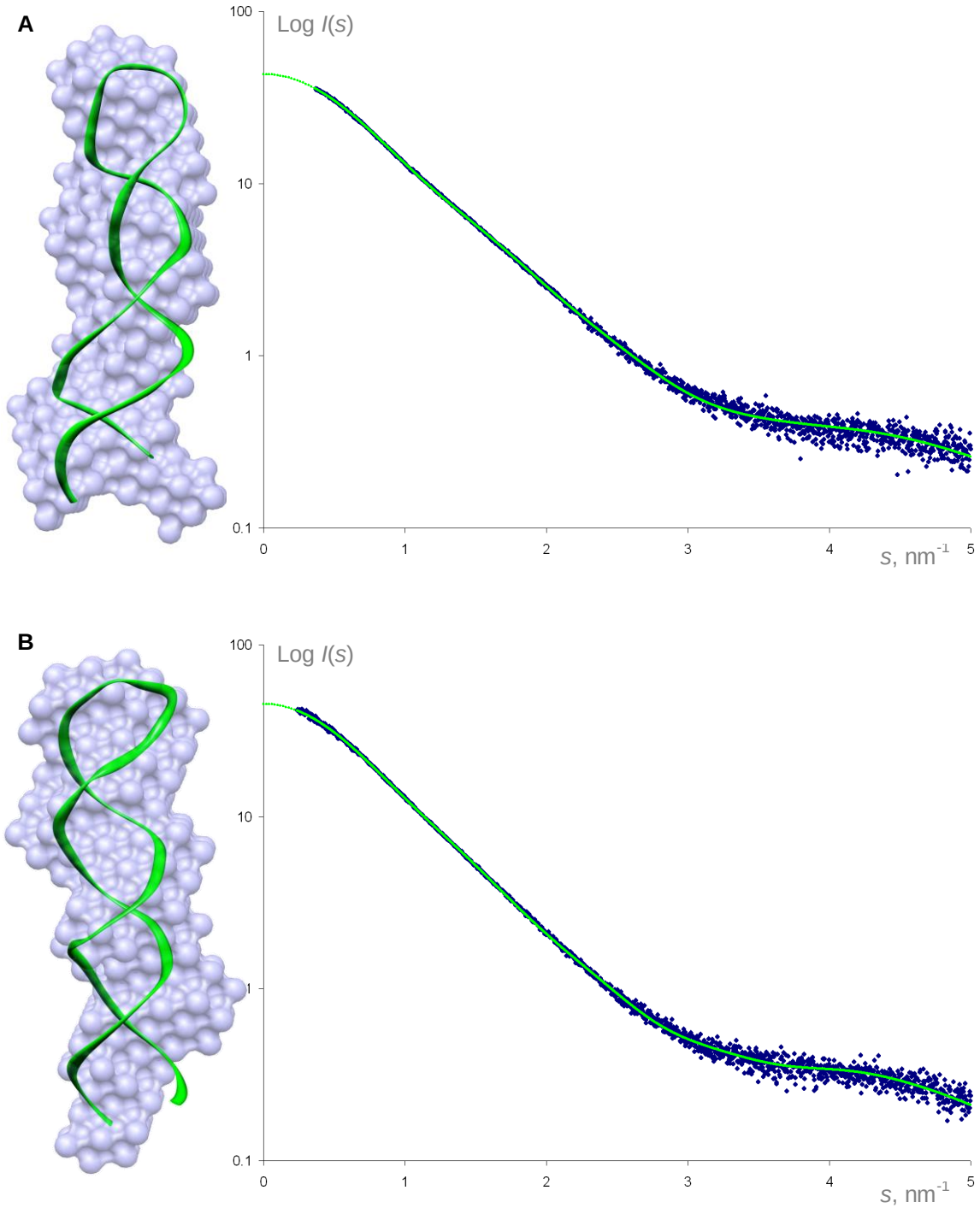


Fig. 5.1 Left: the most probable theoretical models of the NOX E36 (A) and NOX A12 (B) chosen from the output of MC-Fold|MC-Sym using SAXS data as a restriction; overlaid with the *ab initio* DAMMIF models. Right: a comparison of the experimental scattering curve (blue) with the scattering curve calculated from the most probable theoretical model (red) demonstrates a perfect fit.

5.1.2 Chemokines SDF-1 and MCP-1

There are 3D crystal and NMR structures available for both SDF-1 and MCP-1 that can be used to calculate the theoretical scattering with CRYSOL as described above and compared to the experimentally measured scattering. Surprisingly, the SAXS scattering of SDF-1 revealed a monomer in solution, contrast to the crystal structures with the PDB IDs 2NWG, 2J7Z and 1A15 which all showed a dimeric association (Dealwis *et al.* 1998, Murphy *et al.* 2007, Ryu *et al.* 2007). However, the chain B of PDB ID 27JZ fitted neatly the measured scattering curve with a χ value of 1.3.

For MCP-1 two crystal structures of dimers (PDB IDs 1DOK and 3IFD) and one of a monomer (PDB ID 1DOL) in the asymmetric unit were reported (Lubkowski *et al.* 1997). The solution scattering showed a profile in accordance with the dimer described in PDB ID 1DOK although the collected data is of insufficient quality to make a confident conclusion.

	R_g, nm	D_{max}, nm	MM(I_0), kDa	MM(V_{dam}), kDa	V_{Porod}, nm^3
NOX-A12	2.1±0.1	7±1	17±3*	14±1	21±2
NOX-E36	2.1±0.1	7±1	18±3*	13±1	21±2
SDF-1	1.8±0.1	6±1	9±1	13±1	14±2
MCP-1	2.0±0.1	7±1	14±5	13±1	18±2

Table 5.1 Overall parameters of spiegelmers and chemokines determined from experimental data after extrapolation to zero concentration.
*RNAs have approximately two times higher contrast compared to proteins, therefore the I(0) value was divided by two to obtain a correct estimation for the molecular mass.

5.1.3 SDF-1:NOX-A12 and MCP-1:NOX-E36 complexes

The scattering curves of the SDF-1:NOX-A12 complex could not be evaluated due to insufficient quality: the complex disassociated during the measurements depending on the concentration of the samples (Fig. 5.2). Given the high affinity described by NOXXON Pharma AG, this effect was not expected. However, the ratio question to be addressed with this experiment could still be approached for the complex MCP-1:NOX-E36 using preliminary results from rigid body refinement of the complex which was performed using the program SASREF (Petoukhov & Svergun 2005). Given the high resolution models of MCP-1 and NOX-E36, the scattering amplitudes were pre-computed by the program CRYSOL. Starting from a tentative model, this program used simulated annealing to search for a non-overlapping interconnected configuration of chemokines and aptamers fitting the experimental data. Based on the results of a dozen test runs a 2:2 ratio of the two complex components seems most likely; however to obtain an unambiguous structure of the MCP-1:NOX-E36 complex detailed information about the binding epitope is needed.

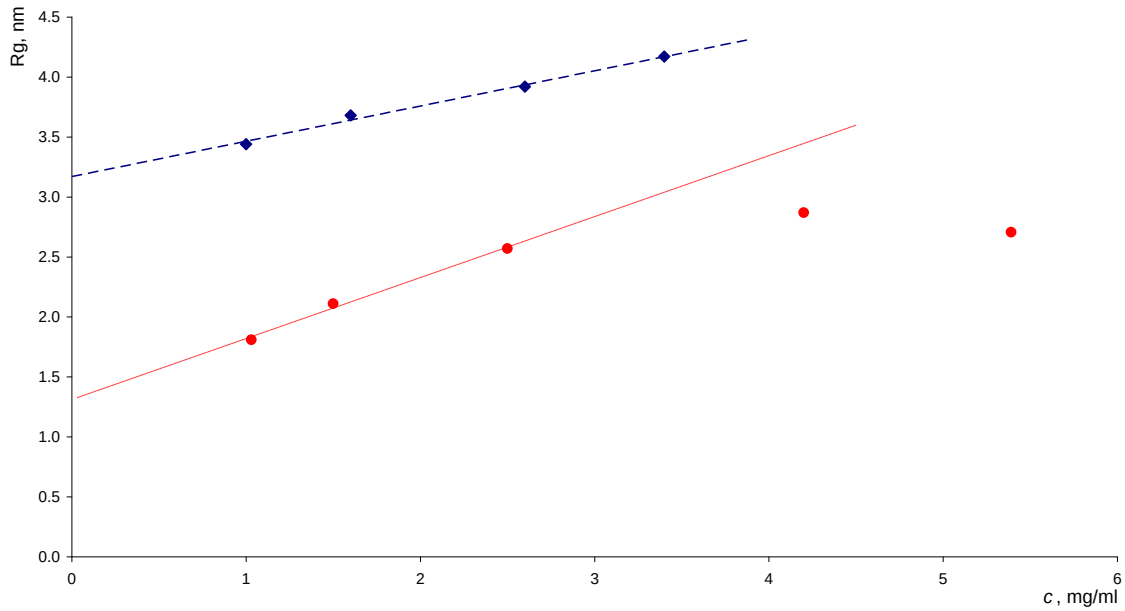


Fig. 5.2 The radius of gyration of the MCP-1:NOX-E36 complex increases linearly with concentration (blue diamonds) which indicates attractive interparticle interactions. In case of the SDF-1:NOX-A12 complex (red circles) such linear dependence is not observed.

5.2 Light-sensitive protein YtvA

Photoreceptors play an important role in plants and, as recently discovered, also in bacteria by converting an extracellular stimulus into an intracellular signal (Losi 2004). One distinct class are blue-light sensitive phototropins harbouring a light, oxygen, voltage (LOV) domain coupled to various effector domains. Photon absorption of the chromophore within the LOV domain results in the activation of the output domain (Crosson *et al.* 2003). The underlying mechanism is the focus of this study.

YtvA from *Bacillus subtilis* is such a light-sensitive photoreceptor comprised of a LOV domain and a sulfate transporter/anti-sigma-factor antagonist (STAS) domain. The protein consists of 261 amino acids and the LOV domain harbours flavin mononucleotide (FMN) as a chromophor (Losi *et al.* 2002). Its simple domain structure renders it very suitable for structural studies and investigation of the activation mechanism.

Upon excitation with blue light (447 nm) a covalent but reversible adduct between the C4a atom of the FMN and Cys62 of the protein forms (Moglich & Moffat 2007). As a result the second domain gets activated and takes part in regulation of sigma-B mediated stress response in *B. subtilis* (Avila-Perez *et al.* 2006). The mechanism of how photon-absorption and adduct-formation in the first domain activates the second domain is not yet understood.

The aim of this study is to investigate by SAXS the oligomeric state of YtvA in solution upon light activation and to determine a low-resolution model of the activated and inactivated protein. The questions addressed are: is the solution structure of YtvA similar to the recently published theoretical model (Avila-Perez *et al.* 2009)? What is the oligomeric state of activated and inactivated YtvA in solution? Is there a conformational change between the activated and inactivated state? How does the effector

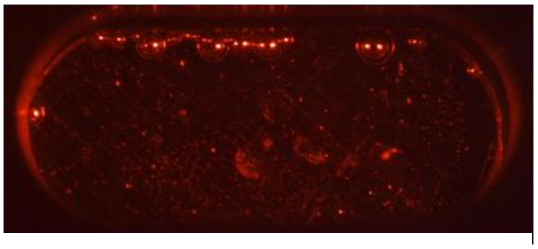


Fig. 5.3 The cell filling control module works even with the red light

activation occur – via dimerization?

To measure the molecular mass and low resolution structure of YtvA upon light activation the sample cell was equipped with red and blue light LEDs that provided the required wavelengths (Fig. 5.4).

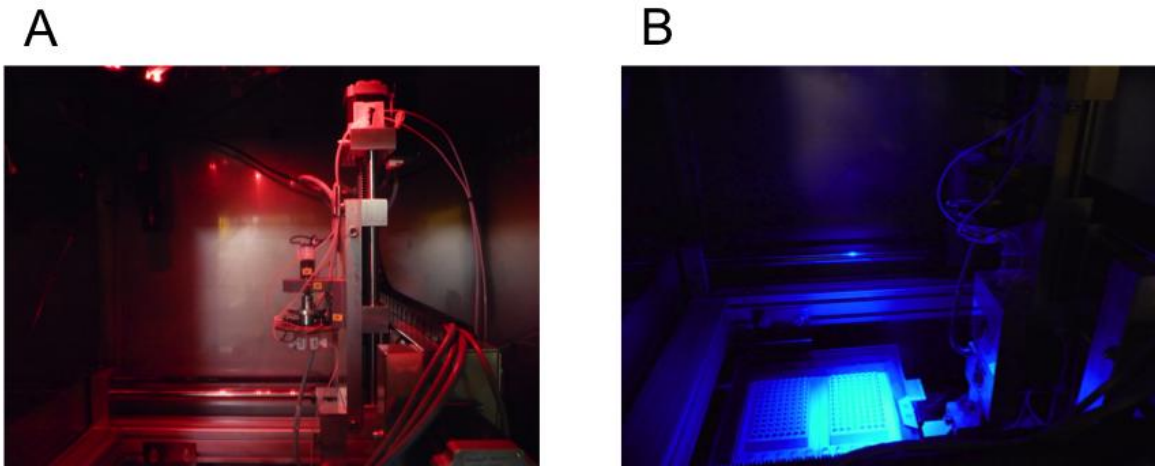


Fig. 5.4 Inside the automated sample changer red light was arranged for inactive YtvA (A) and blue light for active YtvA (B).

5.2.1 Oligomeric state of YtvA upon light activation

Due to changes in light absorption upon activation the concentration of every sample could not be defined with accuracy sufficient to estimate the molecular mass from forward scattering. Comparison of the molecular mass estimated from the dummy atom models yielded a molecular mass of 51±5 kDa and 49±5 kDa for inactive and active YtvA, respectively (Table 5.2). These values correspond to a dimeric state of the protein in solution showing that YtvA does not change its oligomeric state upon light activation. These results are in line with analytical ultracentrifugation (AUC) measurements, which gave a molecular mass of 64.5 kDa for inactivated YtvA, which also suggests a dimeric state of the protein.

	R_g, nm	D_{max}, nm	MM(V_{dam}), kDa	V_p, nm^3
YtvA inactive	3.2±0.1	10±1	51±5	75±8
YtvA active	3.2±0.1	10±1	49±5	74±8
LOV inactive	2.3±0.1	7±1	31±3	46±5
LOV active	2.3±0.3	7±1	32±3	46±5

Table 5.2 Parameters relating to size and shape of active and inactive YtvA determined from scattering profiles after extrapolation to zero concentration.

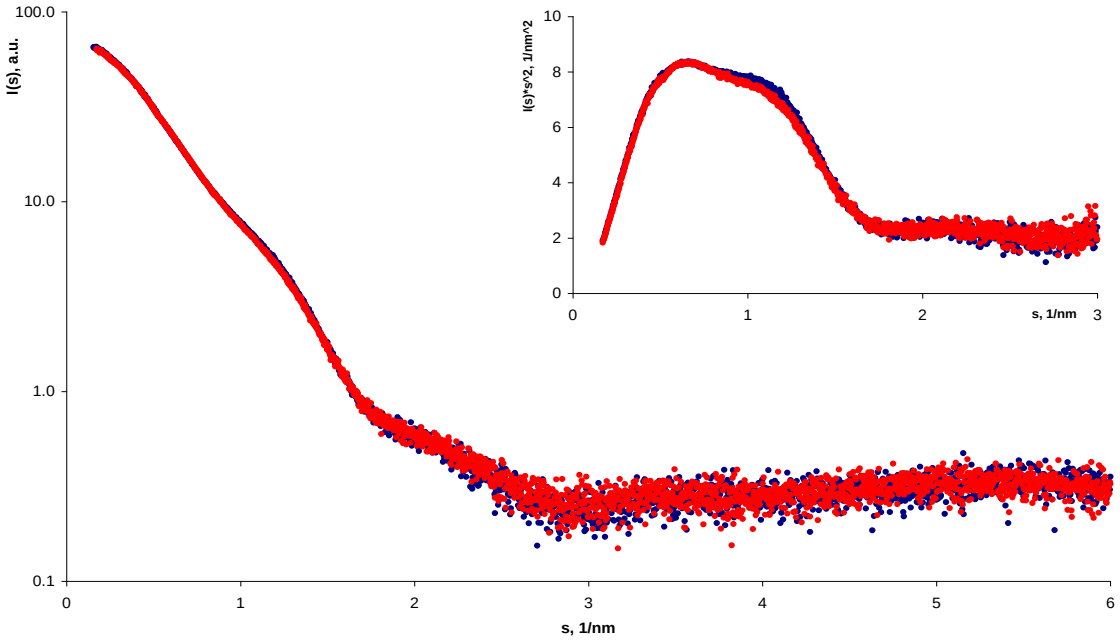


Fig. 5.5 Small-angle x-ray scattering data obtained for inactive (blue) and active (red) YtvA. Curves displayed are derived from an extrapolation to zero concentration of all scattering profiles obtained for six different concentrations in both activation states. Kratky-plot (inset) pronounces angles according to overall shape of the molecule. Difference between active and inactive form of the protein is best visible in the latter plot around 1.0 nm^{-1}

5.2.2 Low resolution models

For the *ab inito* shape determination of the molecules in solution programs DAMMIN (Svergun 1999), DAMMIF (Franke & Svergun 2009) and GASBOR (Svergun *et al.* 2001) were used. All programs produced V-shaped dumbbell models, which fitted very well the experimental data with χ values about 1 (Fig. 5.6A). This indicates that either LOV-LOV-interactions as seen in the high-resolution crystal structure of truncated YtvA (lacking the STAS domain) are necessary to form the dimer of full length YtvA (PDB IDs 2PR5 and 2PR6; (Moglich & Moffat 2007)) or that STAS-STAS-interactions are prone to form the dimer. To further investigate this question we used the program BUNCH (Petoukhov & Svergun 2005) to obtain a most probable rigid body model of dimeric full-length YtvA.

5.2.3 Rigid body model of dimeric YtvA

As a template for the rigid-body modeling the model recently published by the Hellingwerf group was used (Avila-Perez *et al.* 2009). With an average χ^2-value of more than 3 the results for fitting the high-resolution model as a dimer directly to the experimental data was possible, as showed significant deviations were observed against the experimental scattering profiles. The models containing more freedom of rotation for BUNCH resulted in better fits. The three most probable models for inactive YtvA out of the 20 calculated ones are displayed in Fig. 5.6B left together with an overlay of the low-resolution models obtained with DAMMIN/F (Fig. 5.6B middle). BUNCH runs with the rigidity assumed for LOV but not STAS domain, this way leaving reconstruction of the latter to the program, yielded an overall globular shape (Fig. 5.6 right).

An average of 18 structures out of 20 of the calculated models without distance restraints between the two LOV domains within the dimer resulted in the models exhibiting a LOV-LOV-dimer. BUNCH was only able to fit a STAS-STAS-dimer to the experimental solution scattering data with significant worse fits. Altogether, our results strongly support the idea of dimerization mediated by the LOV domains.

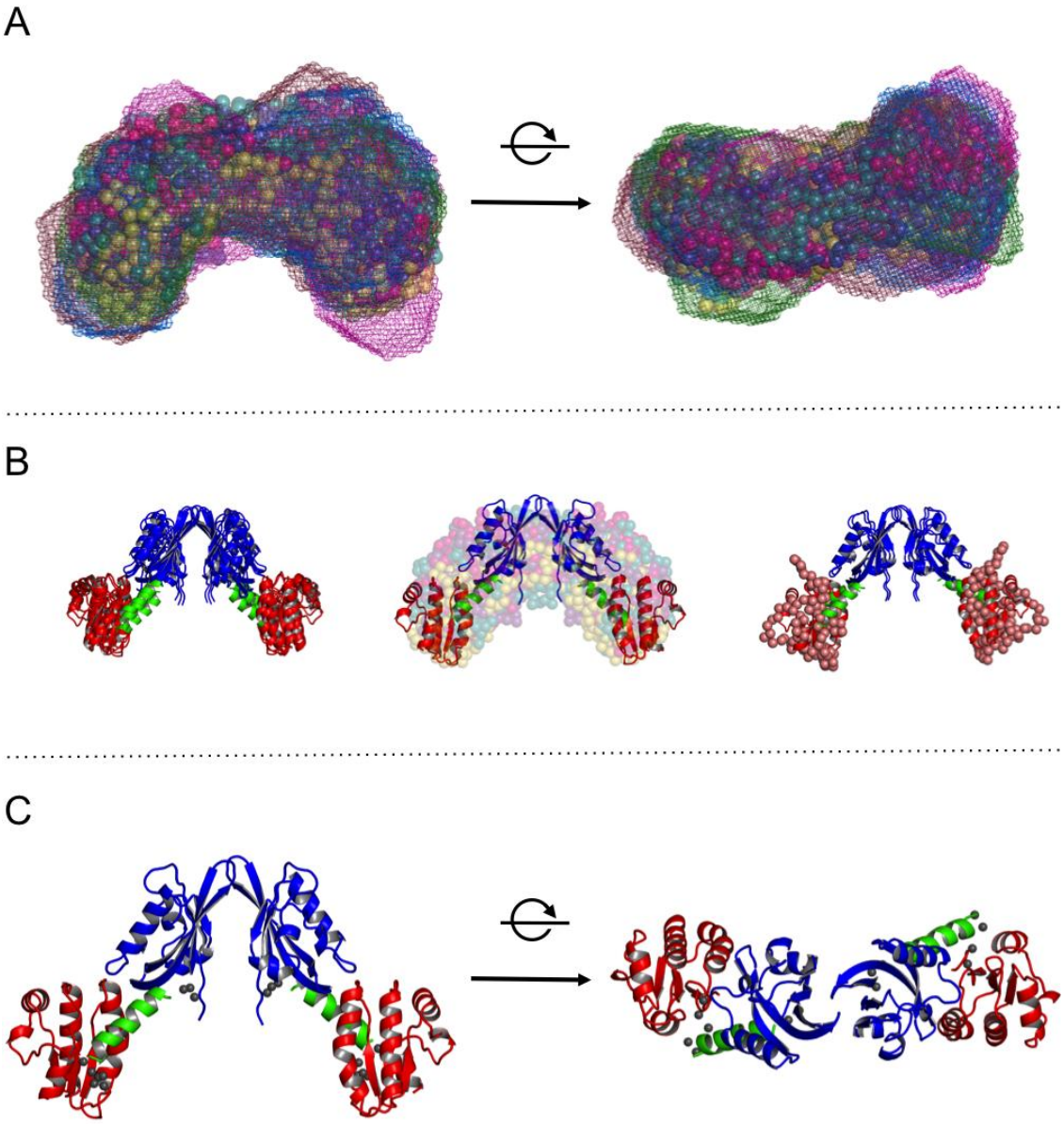


Fig. 5.6 *Ab initio* and rigid body models derived from small-angle x-ray scattering data for YtvA.
A) Overlay of *ab inito* models calculated using GASBOR (spheres) and DAMMIF (mesh). Different colours represent different models. 20 models were calculated with each program and best models were selected using DAMAVER. Four models for each *ab inito* modelling program are displayed. Dimeric YtvA exhibits a V-shaped molecular shape in solution.
B) Rigid body models of YtvA calculated using BUNCH. LOV domain (blue), Jα (green) and STAS (red) are presented. Three out of 20 models correlating best with experimental data are superimposed (left). Superposition of high and low-resolution models displays good correspondence between both methods (middle). BUNCH models were also calculated with and without structural restraints for STAS domain. In the latter case STAS is displayed as spheres with overall globular conformation (right).
C) Best-fit high-resolution model for YtvA using structural restraints derived from a full-length model (Avila-Perez *et al.* 2009) for LOV (residues 21-124), Jα (residues 128-144) and STAS (residues 151-259). Amino acids connecting the three segments were treated as flexible by BUNCH and are represented by grey spheres. For matters of clarity amino acids 1 to 20 are not displayed, though accounted for in BUNCH calculations.

.

These results indicate that YtvA is both in the activated and inactivated state a dimmer, which does not change its overall conformation much upon activation. A similar mechanism was described for other LOV domains from plants (Chen *et al.* 2007, Eitoku *et al.* 2005, Harper *et al.* 2004). Concerning the interactions between the LOV and STAS domains the SAXS data presented in this study showed that firstly, an interaction between LOV and STAS domain is present in the inactive state of the protein and, secondly, that this interaction is not significantly disrupted upon light activation as was shown for *Arabidopsis* LOV2 (Eitoku *et al.* 2007). This finding is in perfect unison with secondary structure estimation between active and inactive YtvA by CD spectroscopy (Buttani *et al.* 2007).

5.3 Dendrimers in methanol

Polyamidoamine (PAMAM) dendrimers are a class of synthetic polymers, which are particularly interesting as efficient cationic polymer vectors for delivering nucleic acids into cells. However, the structural properties of dendrimers in solution have not yet been well described, which would be a significant advance for clinical uses of PAMAM dendrimers as gene vectors. The focus of this study is to determine and compare low-resolution models of two classes of PAMAM dendrimers by SAXS.

RNA interference – gene silencing by double-stranded RNA – and especially the use of *short-interfering* RNA (*si*RNA) in gene therapy hold great promise for the treatment of numerous serious diseases (Hannon 2002), (Sordella *et al.* 2004). To deliver the *si*RNA into the cell a vector is required to cross the biological barrier of the cell membrane. During the past few years PAMAM dendrimers have been investigated as *si*RNA delivery systems due to their promising structural and biological properties (Roberts *et al.* 1996), (Esfand & Tomalia 2001). In contrast to classical linear chain polymers dendrimers are branched in a well defined three-dimensional structure and are usually polyvalent, monodisperse and biocompatible with a low toxicity (Mourey *et al.* 1992), (Jevprasesphant *et al.* 2003).

In this study two classes of PAMAM dendrimers with ethylenediamine (EDA) and triethanolamine (TEA) cores in methanol were investigated. This solvent ensures

dendrimer stability but is hard to handle with the sample changer. Thanks to the cell filling control software it was possible to avoid manual filling and minimize failures.

For each EDA PAMAM dendrimer extension generation (generation four (G4) to generation seven (G7)) several concentrations in the range 0.1 - 5 % wt/vol (up to 10 % for G4 EDA) were measured. For TEA PAMAM dendrimers (G4 and G5) only two concentrations were measured (0.1 and 1 % wt/vol) due to limited availability of the synthesized dendrimers. SAXS patterns for EDA PAMAM dendrimers from G4 to G7 (Fig. 5.7) are in accordance with data by (Prosa *et al.* 2001) showing a change in shape as a function of the dendrimer generation. The SAXS patterns for TEA dendrimers G4 and G5 show a more elongated shape compared to equivalent EDA generations.

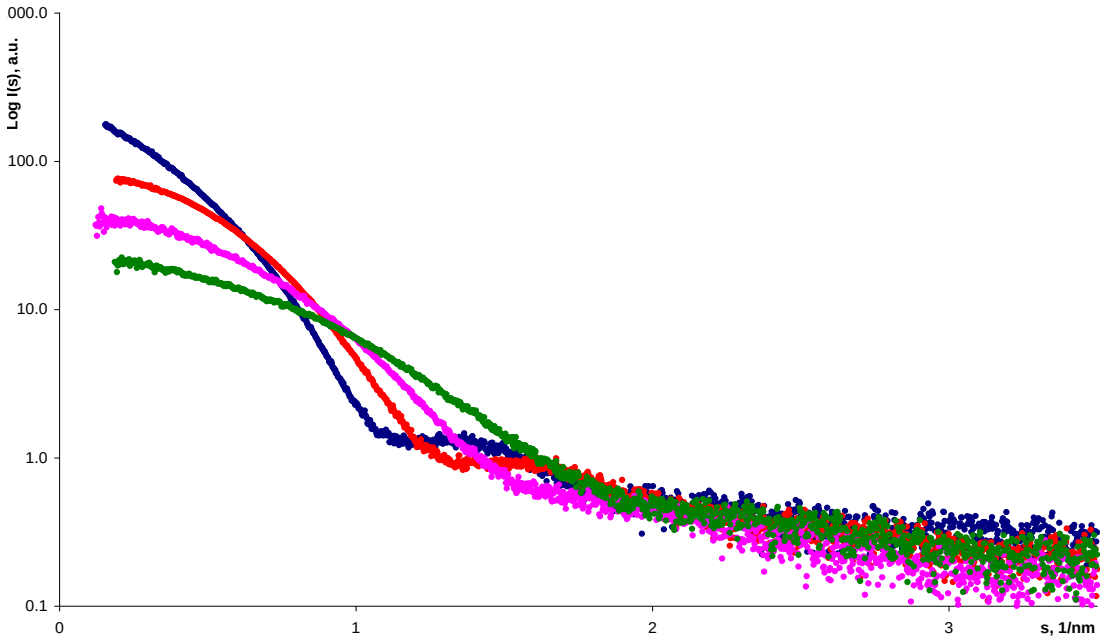


Fig. 5.7 SAXS patterns of EDA PAMAM dendrimers: G4 (green), G5 (magenta), G6 (red), G7 (blue).

The data obtained from the SAXS measurements are summarized in Table 5.3. Due to high volatility of methanol the concentration of every sample could not be defined with accuracy sufficient to estimate the molecular mass from forward scattering, therefore it is omitted. As expected the D_{max} of EDA PAMAM increases as a function of generation, from G4 to G7. Also R_g and the excluded volume of the hydrated dendrimers (Porod volume, V_p) increase as a function of the generation, linearly for R_g and D_{max} and with the power of 3 for V_P. A comparison of the V_P calculated from the experimental data to the calculated DAMMIF volume from modeling for EDA and TEA PAMAM dendrimers for G4 and G5 shows a quite well correlation.

Although TEA dendrimers are smaller in molecular mass and the surface group number as compared to EDA PAMAMs with an equivalent generation, TEA PAMAM dendrimers show a larger and more elongated structure. This could be attributed to a larger solvation and hydration of the TEA molecules.

The data show that EDA and TEA dendrimers in methanol are mainly monomeric. The molecular masses for both dendrimers in methanol calculated from the experimental data are larger than expected from the theoretical structures, which is probably due to errors of the dendrimer concentrations in methanol.

	R_g, nm	D_{max}, nm	V_{Porod}, nm^3
EDA G4	2.0±0.1	6±1	39±4
EDA G5	2.4±0.1	8±1	77±8
EDA G6	2.9±0.1	10±1	143±14
EDA G7	4.1±0.1	13±1	294±29
TEA G4	2.4±0.1	8±1	40±4
TEA G5	3.9±0.1	13±1	109±11

Table 5.3 Overall Parameters of EDA and TEA dendrimers determined from experimental data after extrapolation to zero concentration.

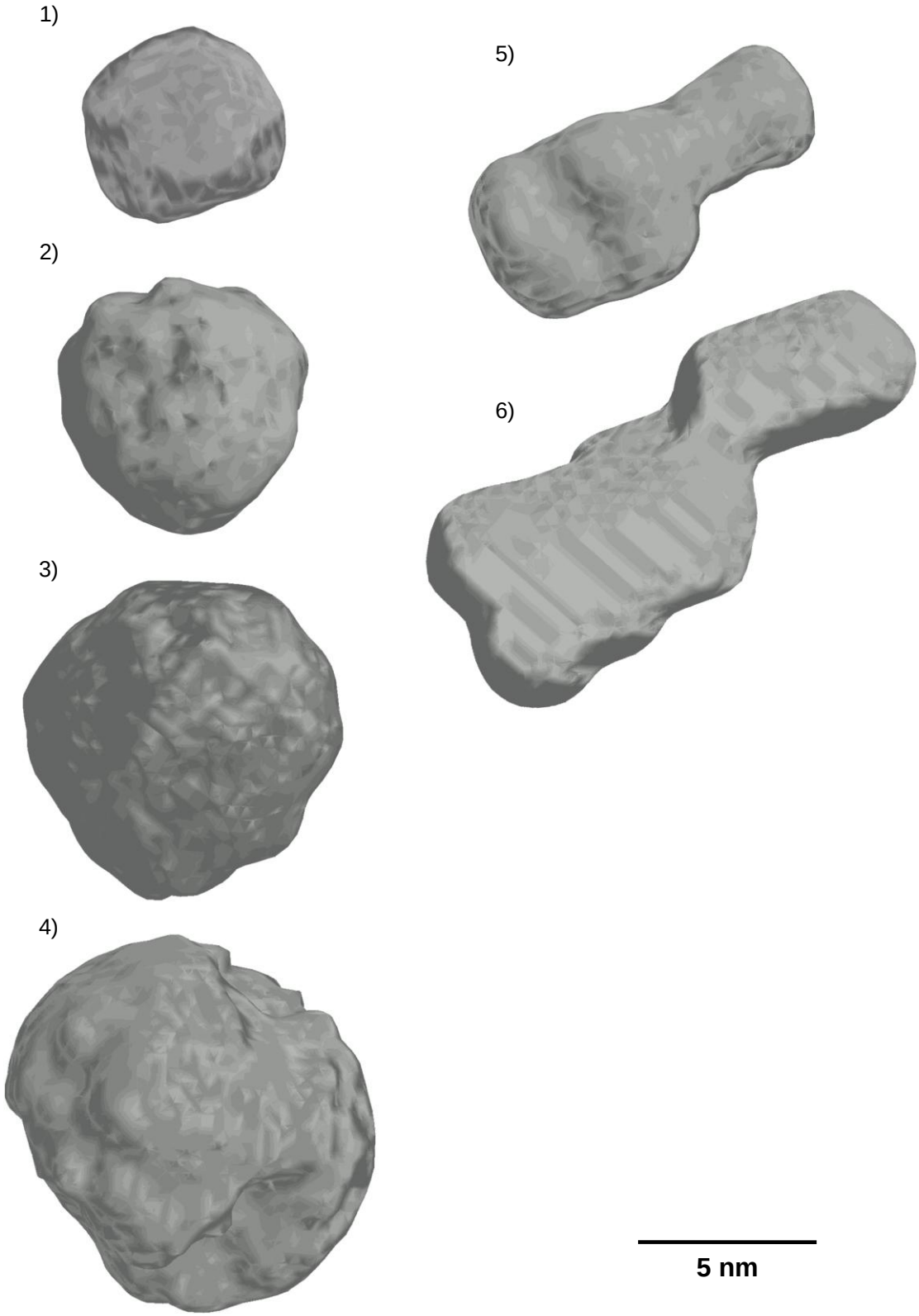


Fig. 5.8 *Ab initio* models of dendrimers: 1) EDA G4; 2) EDA G5; 3) EDA G6; 4) EDA G7; 5) TEA G4; 6) TEA G5.

5.4 Discussion

The new sample cell visualization setup coupled with the image recognition module dramatically reduced the number of failures during the measurements, particularly in the case of particles in methanol. Thanks to the possibility of changing the LEDs from white to red it was possible to measure light-sensitive samples. The automated data processing pipeline successfully evaluated data collected from samples of different nature (proteins, nucleic acids, dendrimers) immediately creating a summary of the overall parameters which helped to make important conclusions about the quality of the samples and their oligomeric states. The generated *ab initio* models were used as a starting point for further modeling.

Conclusions

In the present study, acute problems for full automation of SAXS measurements are addressed. In particular, a comprehensive solution for rapid data collection, processing, analysis and storage is offered, making SAXS a viable method for high throughput studies.

An improved system for sample monitoring has been developed with a new design of sample cell visualization hardware allowing automated control of the cell filling process. An image processing software module detects improper cell filling and attempts to correct it providing feedback to the sample changer robot. This setup is implemented at the X33 beamline of EMBL Hamburg and is used together with two robotic sample changers.

A number of data processing tools were developed to make automated data analysis possible. A Guinier approximation based algorithm reliably estimates the radius of gyration and experimental data quality providing valuable information that is necessary for other data reduction tools. Data collected from multiple concentrations is examined and automatically extrapolated to infinite dilution; various overall parameters are calculated from the data.

A high-throughput data analysis pipeline was developed, performing data evaluation in real time during the experiment without user intervention. Sequential and parallel data processing using multiple software modules, both existing and newly developed, is complemented with cross-validation of intermediate results to ensure a robust reproducible outcome. The pipeline includes intelligent background subtraction, analysis of concentration effects, calculation of overall parameters and characteristic functions, low resolution *ab initio* shape reconstruction and the use of a SAXS patterns database. The intermediate and final results are stored in a structured user-friendly web-compatible XML format.

The feedback from over 100 external user groups testing the automated pipeline during 2009 indicates that an automatically generated overview of the overall parameters saves valuable time and greatly improves the reliability of the results; the immediate feedback during the measurements allows users to adjust the experimental

conditions “on-the-fly” if necessary. The data analysis pipeline is ready for installation at high brilliance SAXS beamlines of the third generation synchrotron radiation sources, e.g. the BioSAXS beamline of Petra-III; currently the pipeline is tested at the ESRF beamline ID14-3.

The flexible modular design of the automated SAXS pipeline allows for the introduction of further decision making blocks and the extensive use of *a priori* information including e.g. protein/nucleic acid sequences, available high-resolution models, and neutron scattering data. These developments provide a foundation for an intelligent fully automatic expert system for SAS-based model building.

References

1. Avila-Perez, M., Hellingwerf, K. J., and Kort, R. (2006) Blue light activates the sigmaB-dependent stress response of Bacillus subtilis via YtvA *J Bacteriol* **188**, 6411-4.

2. Avila-Perez, M., Vreede, J., Tang, Y., Bende, O., Losi, A., Gartner, W., and Hellingwerf, K. (2009) In vivo mutational analysis of YtvA from Bacillus subtilis: mechanism of light activation of the general stress response *J Biol Chem* **284**, 24958-64.

3. Bartkiewicz, P., and Duval, P. (2007) TINE as an accelerator control system at DESY *Measurement Science & Technology* **18**, 2379-2386.

4. Bernado, P., Mylonas, E., Petoukhov, M. V., Blackledge, M., and Svergun, D. I. (2007) Structural characterization of flexible proteins using small-angle X-ray scattering *J Am Chem Soc* **129**, 5656-64.

5. Broxmeyer, H. E. (2008) Chemokines in hematopoiesis *Curr Opin Hematol* **15**, 49-58.

6. Buttani, V., Losi, A., Eggert, T., Krauss, U., Jaeger, K. E., Cao, Z., and Gartner, W. (2007) Conformational analysis of the blue-light sensing protein YtvA reveals a competitive interface for LOV-LOV dimerization and interdomain interactions *Photochemical & Photobiological Sciences* **6**, 41-49.

7. Chen, E., Swartz, T. E., Bogomolni, R. A., and Kliger, D. S. (2007) A LOV story: the signaling state of the phot1 LOV2 photocycle involves chromophore-triggered protein structure relaxation, as probed by far-UV time-resolved optical rotatory dispersion spectroscopy *Biochemistry* **46**, 4619-24.

8. Crosson, S., Rajagopal, S., and Moffat, K. (2003) The LOV domain family: photoresponsive signaling modules coupled to diverse output domains *Biochemistry* **42**, 2-10.

9. David, G., and Perez, J. (2009) Combined sampler robot and high-performance liquid chromatography: a fully automated system for biological small-angle X-ray scattering experiments at the Synchrotron SOLEIL SWING beamline *Journal of Applied Crystallography* **42**, 892-900.

10. Dealwis, C., Fernandez, E. J., Thompson, D. A., Simon, R. J., Siani, M. A., and Lolis, E. (1998) Crystal structure of chemically synthesized [N33A] stromal cell-derived factor 1alpha, a potent ligand for the HIV-1 "fusin" coreceptor *Proc Natl Acad Sci U S A* **95**, 6941-6.

11. Eitoku, T., Nakasone, Y., Matsuoka, D., Tokutomi, S., and Terazima, M. (2005) Conformational dynamics of phototropin 2 LOV2 domain with the linker upon photoexcitation *Journal of the American Chemical Society* **127**, 13238-13244.

12. Eitoku, T., Nakasone, Y., Zikihara, K., Matsuoka, D., Tokutomi, S., and Terazima, M. (2007) Photochemical intermediates of Arabidopsis phototropin 2 LOV domains associated with conformational changes *Journal of Molecular Biology* **371**, 1290-1303.

13. Esfand, R., and Tomalia, D. A. (2001) Poly(amidoamine) (PAMAM) dendrimers: from biomimicry to drug delivery and biomedical applications *Drug Discovery Today* **6**, 427-436.

14. Esposito, C., Petoukhov, M. V., Svergun, D. I., Ruggiero, A., Pedone, C., Pedone, E., and Berisio, R. (2008) Evidence for an elongated dimeric structure of heparin-binding hemagglutinin from Mycobacterium tuberculosis *J Bacteriol* **190**, 4749-53.

15. Franke, D., and Svergun, D. I. (2009) DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering *J. Appl. Cryst.* **42**, 342-346.

16. Funari, S. S., Rapp, G., Perbandt, M., Dierks, K., Vallazza, M., Betzel, C., Erdmann, V. A., and Svergun, D. I. (2000) Structure of free thermus flavus 5 S rRNA at 1.3 nm resolution from synchrotron X-ray solution scattering *J Biol Chem* **275**, 31283-8.

17. Gabel, F., Simon, B., Nilges, M., Petoukhov, M., Svergun, D., and Sattler, M. (2008) A structure refinement protocol combining NMR residual dipolar couplings and small angle scattering restraints *Journal of Biomolecular Nmr* **41**, 199-208.

18. Glatter, O. (1977) A new method for the evaluation of small-angle scattering data *J. Appl. Cryst.* **10**, 415-421.

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

20. Guinier, A. (1939) La diffraction des rayons X aux tres petits angles; application a l'etude de phenomenes ultramicroscopiques *Ann. Phys. (Paris)* **12**, 161-237.

21. Hannon, G. J. (2002) RNA interference *Nature* **418**, 244-251.

22. Harper, S. M., Neil, L. C., Day, I. J., Hore, P. J., and Gardner, K. H. (2004) Conformational changes in a photosensory LOV domain monitored by time-resolved NMR spectroscopy *Journal of the American Chemical Society* **126**, 3390-3391.

23. Hura, G. L., Menon, A. L., Hammel, M., Rambo, R. P., Poole, F. L., 2nd, Tsutakawa, S. E., Jenney, F. E., Jr., Classen, S., Frankel, K. A., Hopkins, R. C., Yang, S. J., Scott, J. W., Dillard, B. D., Adams, M. W., and Tainer, J. A. (2009) Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS) *Nat Methods* **6**, 606-12.

24. Ilavsky, J., and Jemian, P. R. (2009) Irena: tool suite for modeling and analysis of small-angle scattering *Journal of Applied Crystallography* **42**, 347-353.

25. Jevprasesphant, R., Penny, J., Jalal, R., Attwood, D., McKeown, N. B., and D'Emanuele, A. (2003) The influence of surface modification on the cytotoxicity of PAMAM dendrimers *International Journal of Pharmaceutics* **252**, 263-266.

26. Koch, M. H., Vachette, P., and Svergun, D. I. (2003) Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution *Q Rev Biophys* **36**, 147-227.

27. Koch, M. H. J., Vachette, P., and Svergun, D. I. (2003) Small angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution *Quart. Rev. Biophys.* **36**, 147-227.

28. Konarev, P. V., Petoukhov, M. V., Volkov, V. V., and Svergun, D. I. (2006) ATSAS 2.1, a program package for small-angle scattering data analysis *J. Appl. Crystallogr.* **39**, 277-286.

29. Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J., and Svergun, D. I. (2003) PRIMUS - a Windows-PC based system for small-angle scattering data analysis *J. Appl. Crystallogr.* **36**, 1277-1282.

30. Kozin, M. B., and Svergun, D. I. (2001) Automated matching of high- and low-resolution structural models *J. Appl. Crystallogr.* **34**, 33-41.

31. Krissinel, E., and Henrick, K. (2007) Inference of macromolecular assemblies from crystalline state *J Mol Biol* **372**, 774-97.

32. Losi, A. (2004) The bacterial counterparts of plant phototropins *Photochem Photobiol Sci* **3**, 566-74.

33. Losi, A., Polverini, E., Quest, B., and Gartner, W. (2002) First evidence for phototropin-related blue-light receptors in prokaryotes *Biophys J* **82**, 2627-34.

34. Lubkowski, J., Bujacz, G., Boque, L., Domaille, P. J., Handel, T. M., and Wlodawer, A. (1997) The structure of MCP-1 in two crystal forms provides a rare example of variable quaternary interactions *Nat Struct Biol* **4**, 64-9.

35. Maasch, C., Buchner, K., Eulberg, D., Vonhoff, S., and Klussmann, S. (2008) Physicochemical stability of NOX-E36, a 40mer L-RNA (Spiegelmer) for therapeutic applications *Nucleic Acids Symp Ser (Oxf)*, 61-2.

36. Maddison, D. R., Swofford, D. L., and Maddison, W. P. (1997) NEXUS: an extensible file format for systematic information *Syst Biol* **46**, 590-621.

37. Malfois, M., and Svergun, D. I. (2000) SasCIF - an extension of core Crystallographic Information File for small angle scattering *J. Appl. Crystallogr.* **34**, 812-816.

38. McPhillips, T. M., McPhillips, S. E., Chiu, H. J., Cohen, A. E., Deacon, A. M., Ellis, P. J., Garman, E., Gonzalez, A., Sauter, N. K., Phizackerley, R. P., Soltis, S. M., and Kuhn, P. (2002) Blu-Ice and the Distributed Control System: software for data acquisition and instrument control at macromolecular crystallography beamlines *J Synchrotron Radiat* **9**, 401-6.

39. Moglich, A., and Moffat, K. (2007) Structural basis for light-dependent signaling in the dimeric LOV domain of the photosensor YtvA *J Mol Biol* **373**, 112-26.

40. Mourey, T. H., Turner, S. R., Rubinstein, M., Frechet, J. M. J., Hawker, C. J., and Wooley, K. L. (1992) Unique Behavior of Dendritic Macromolecules - Intrinsic-Viscosity of Polyether Dendrimers *Macromolecules* **25**, 2401-2406.

41. Murphy, J. W., Cho, Y., Sachpatzidis, A., Fan, C., Hodsdon, M. E., and Lolis, E. (2007) Structural and functional basis of CXCL12 (stromal cell-derived factor-1 alpha) binding to heparin *J Biol Chem* **282**, 10018-27.

42. Mylonas, E., and Svergun, D. I. (2007) Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering *J. Appl. Cryst.* **40**, s245-s249.

43. Parisien, M., and Major, F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data *Nature* **452**, 51-5.

44. Petoukhov, M. V., Konarev, P. V., Kikhney, A. G., and Svergun, D. I. (2007) ATSAS 2.1 - towards automated and web-supported small-angle scattering data analysis *J. Appl. Cryst.* **40**, s223-s228.

45. Petoukhov, M. V., and Svergun, D. I. (2005) Global rigid body modeling of macromolecular complexes against small-angle scattering data *Biophys J* **89**, 1237-50.

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

47. Petoukhov, M. V., and Svergun, D. I. (2007) Analysis of X-ray and neutron scattering from biomacromolecular solutions *Curr Opin Struct Biol* **17**, 562-571.

48. Porod, G. (1951) Die Röntgenkleinwinkelstreuung von dichtgepackten kolloiden Systemen, I *Kolloid-Z.* **124**, 83-114.

49. Porod, G. (1982) in *Small-angle X-ray scattering* (Glatter, O., and Kratky, O., Eds.) pp 17-51, Academic Press, London.

50. Prosa, T. J., Bauer, B. J., and Amis, E. J. (2001) From stars to spheres: A SAXS analysis of dilute dendrimer solutions *Macromolecules* **34**, 4897-4906.

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

52. Roberts, J. C., Bhalgat, M. K., and Zera, R. T. (1996) Preliminary biological evaluation of polyamidoamine (PAMAM) Starburst(TM) dendrimers *Journal of Biomedical Materials Research* **30**, 53-65.

53. Roessle, M. W., Klaering, R., Ristau, U., Robrahn, B., Jahn, D., Gehrmann, T., Konarev, P., Round, A., Fiedler, S., Hermes, C., and Svergun, D. (2007) Upgrade of the small-angle X-ray scattering beamline X33 at the European Molecular Biology Laboratory, Hamburg *J. Appl. Cryst.* **40**, s190-s194.

54. Rolbin, Y. A., Kayushina, R. L., Feigin, L. A., and Schedrin, B. M. (1973) Computer calculations of the X-ray small-angle scattering by macromolecule models *Kristallografia (in Russian)* **18**, 701-705.

55. Round, A. R., Franke, D., Moritz, S., Huchler, R., Fritsche, M., Malthan, D., Klaering, R., Svergun, D. I., and Roessle, M. (2008) Automated sample-changing robot for solution scattering experiments at the EMBL Hamburg SAXS station X33 *Journal of Applied Crystallography* **41**, 913-917.

56. Ruster, C., and Wolf, G. (2008) The role of chemokines and chemokine receptors in diabetic nephropathy *Front Biosci* **13**, 944-55.

57. Ryu, E. K., Kim, T. G., Kwon, T. H., Jung, I. D., Ryu, D., Park, Y. M., Kim, J., Ahn, K. H., and Ban, C. (2007) Crystal structure of recombinant human stromal cell-derived factor-1alpha *Proteins* **67**, 1193-7.

58. Sayyed, S. G., Hagele, H., Kulkarni, O. P., Endlich, K., Segerer, S., Eulberg, D., Klussmann, S., and Anders, H. J. (2009) Podocytes produce homeostatic chemokine stromal cell-derived factor-1/CXCL12, which contributes to glomerulosclerosis, podocyte loss and albuminuria in a mouse model of type 2 diabetes *Diabetologia* **52**, 2445-54.

59. Semenyuk, A. V., and Svergun, D. I. (1991) GNOM - a program package for small-angle scattering data processing *J. Appl. Crystallogr.* **24**, 537-540.

60. Sokolova, A. V., Volkov, V. V., and Svergun, D. I. (2003) Prototype of database for rapid protein classification based on solution scattering data *J. Appl. Crystallogr.* **36**, 865-868.

61. Sordella, R., Bell, D. W., Haber, D. A., and Settleman, J. (2004) Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways *Science* **305**, 1163-1167.

62. Steel, R. G. D., and Torrie, J. H. (1960) *Principles and procedures of statistics, with special reference to the biological sciences*, McGraw-Hill, New York,.

63. Svergun, D. I. (1992) Determination of the regularization parameter in indirect-transform methods using perceptual criteria *J. Appl. Crystallogr.* **25**, 495-503.

64. Svergun, D. I. (1999) Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing *Biophys J* **76**, 2879-86.

65. Svergun, D. I., Barberato, C., and Koch, M. H. J. (1995) CRYSOL - a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates *J. Appl. Crystallogr.* **28**, 768-773.

66. Svergun, D. I., and Nierhaus, K. H. (2000) A map of protein-rRNA distribution in the 70 S Escherichia coli ribosome *J Biol Chem* **275**, 14432-9.

67. Svergun, D. I., Petoukhov, M. V., and Koch, M. H. (2001) Determination of domain structure of proteins from X-ray solution scattering *Biophys J* **80**, 2946-53.

68. Svergun, D. I., Petoukhov, M. V., and Koch, M. H. J. (2001) Determination of domain structure of proteins from X-ray solution scattering *Biophys J* **80**, 2946-53.

69. Svergun, D. I., Richard, S., Koch, M. H., Sayers, Z., Kuprin, S., and Zaccai, G. (1998) Protein hydration in solution: experimental observation by x-ray and neutron scattering *Proc Natl Acad Sci U S A* **95**, 2267-72.

70. Tardieu, A. (1994) in *Neutron and Synchrotron Radiation For Condensed Matter Studies* pp 145-160, Les editions de Physique (France) Springer-Verlag.

71. Toft, K. N., Vestergaard, B., Nielsen, S. S., Snakenborg, D., Jeppesen, M. G., Jacobsen, J. K., Arleth, L., and Kutter, J. P. (2008) High-throughput Small Angle X-ray Scattering from proteins in solution using a microfluidic front-end *Analytical Chemistry* **80**, 3648-3654.

72. Tucci, M., Ciavarella, S., Strippoli, S., Dammacco, F., and Silvestris, F. (2009) Oversecretion of cytokines and chemokines in lupus nephritis is regulated by intraparenchymal dendritic cells: a review *Ann N Y Acad Sci* **1173**, 449-57.

73. Vérétout, F., Delaye, M., and Tardieu, A. (1989) Molecular basis of eye lens transparency. Osmotic pressure and X-ray analysis of α-crystallin solutions. *J. molec. Biol.* **205**, 713-728.

74. Volkov, V. V., and Svergun, D. I. (2003) Uniqueness of ab initio shape determination in small angle scattering *J. Appl. Crystallogr.* **36**, 860-864.

75. Winn, M. D., Ashton, A. W., Briggs, P. J., Ballard, C. C., and Patel, P. (2002) Ongoing developments in CCP4 for high-throughput structure determination *Acta Crystallogr D Biol Crystallogr* **58**, 1929-36.

76. Zheng, W., Johnston, S. A., and Joshua-Tor, L. (1998) The unusual active site of Gal6/bleomycin hydrolase can act as a carboxypeptidase, aminopeptidase, and peptide ligase *Cell* **93**, 103-9.

77. http://qtsoftware.com

78. http://www.w3c.org/xml

Acknowledgements

First of all I would like to thank Maxim Petoukhov for introducing me to the world of small angle scattering. Then I would like to thank Daniel Franke for motivating me to improve my programming style and helping me with coding issues. I thank Stratos Mylonas for guiding me through the first steps of my PhD, Christian Gorba and Haydyn Mertens for proofreading my thesis, Manfred Roessle for countenance and the rest of the BioSAXS group of EMBL Hamburg for the great working atmosphere. Many thanks to the DESY and EMBL administration for making my life easier. I acknowledge SAXIER DS8 for the financial support.

In this work Maxim Petoukhov and Peter Konarev contributed developing the DATTOOLS; Daniel Franke developed the BMS. I would like to gratefully acknowledge my collaboration partners: Bara Schmidt (section 5.1 Protein-RNA complexes: chemokines inhibited by spiegelmers), Marcel Jurk (section 5.2 Light-sensitive protein YtvA) and Louiza Zerrad (section 5.3 Dendrimers in methanol).

I thank my parents Lyubov and Gennady Kikhney for encouraging me to do science and my bride Judith Schmiedel for motivating me to complete things that I started once.

I would like to thank my university supervisor Christian Betzel and my Thesis Advisory Committee members Andreas Ladurner and Christoph Hermes for supporting my work. My greatest thanks go to Dmitri Svergun for providing me with exiting and challenging tasks and being an excellent supervisor.

Appendix

Curriculum vitae

Personal data

Name:	Alexey Kikhney
Address:	EMBL c/o DESY, Notkestrasse 85, Geb. 25 A, 22607 Hamburg Tel.: +49 40 89902-170 plmnnk@embl-hamburg.de
Date of birth:	27.03.1979
Birth-place:	Primorsky Krai, USSR
Nationality:	Russian

Education

Diploma in Computer Science, June 2001
Lomonosov Moscow State University, Moscow, Russian Federation

Related experience

Pre-Doctoral Fellow, European Molecular Biology Laboratory, Hamburg Outstation, March 2006 - present

Software development for high-throughput biological small angle X-ray scattering.

Project Manager, IT4profit Ltd, Moscow office, October 2002 – February 2006

High-intensity internet applications development.
Cross-platform data migration.

Research Assistant, Shirshov Institute of Oceanology RAS, October 2001 – May 2003

Development of an integrated system for biological, geological and physical data collection with a precise navigation binding.

Development and implementation of a transponder calibrating system as a part of long-base acoustic navigation system for *Mir* submersibles.

Web-developer, IT4profit Ltd, Moscow office, May 2001 – April 2002

Web-programming (XSLT, JSP, SQL), interface design, usability analysis.

Student, Lomonosov Moscow State University, Computational Mathematics and Cybernetics department, chair of System Programming, September 1996 – June 2001

Diploma: Syntax-directed Flow Chart Editor Development. Design and implementation of a part of syntax-directed editing software package – a tool for visual software development.

Publications

Peer reviewed publication

Petoukhov, M. V., Konarev, P. V., Kikhney, A. G., and Svergun, D. I. (2007) ATSAS 2.1 - towards automated and web-supported small-angle scattering data analysis *J. Appl. Cryst.* **40**, s223-s228.

Conference proceedings

Kikhney, A.G., Franke, D., Konarev, P.V., Gajda, M.J., Petoukhov, M. V. and Svergun D.I. (2009) Automated small-angle X-ray scattering data processing and analysis pipeline. *XIV International Conference on Small-Angle Scattering, Oxford, UK* (poster).

Kikhney, A.G., Franke, D., Konarev, P.V., Gajda, M.J., Petoukhov, M. V. and Svergun D.I. (2009) Automated small-angle X-ray scattering data processing and analysis pipeline. *First International Symposium on Structural Systems Biology, Hamburg, Germany* (poster).

Kikhney, A.G., Franke, D., Konarev, P.V. and Svergun D.I. (2008) Software for automated high-throughput biological small-angle X-ray scattering *Acta Cryst.* **A64**, C554-555. *XXI Congress and General Assembly of the International Union of Crystallography, Osaka, Japan* (poster).

Kikhney, A.G., Konarev, P.V. and Svergun D.I. (2007) Software for automated high-throughput biological small-angle X-ray scattering. *9th International Conference on Biology and Synchrotron Radiation, Manchester, UK* (award-winning poster).

Eidesstattliche Erklärung

Hiermit versichere ich eidesstattlich, dass ich die vorliegende Arbeit selbständig und ohne unerlaubte fremde Hilfe angefertigt habe. Ich habe keine anderen als die im Literaturverzeichnis angeführten Quellen benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Zusammenarbeiten sind im Text genannt.

Die Arbeit wurde zuvor keiner Prüfungsbehörde in gleicher oder ähnlicher Form vorgelegt.

Alexey Kikhney
Hamburg, den 29. April 2010