BIOLOGICAL CRYSTALLOGRAPHY

ISSN 1399-0047

DATASW, a tool for HPLC–SAXS data analysis

Alexander V. Shkumatov* and Sergei V. Strelkov

Department of Pharmaceutical and Pharmacological Sciences, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. *Correspondence e-mail: alexander.shkumatov@gmail.com

Received 27 January 2015
Accepted 9 April 2015

Edited by Z. S. Derewenda, University of Virginia, USA

Keywords: SAXS; HPLC; oligomer; processing.

Supporting information: this article has supporting information at journals.iucr.org/d

Small-angle X-ray scattering (SAXS) in solution is a common low-resolution method which can efficiently complement the high-resolution information obtained by crystallography or NMR. Sample monodispersity is key to reliable SAXS data interpretation and model building. Beamline setups with inline high-performance liquid chromatography (HPLC) are particularly useful for accurate profiling of heterogeneous samples. The program *DATASW* performs averaging of individual data frames from HPLC–SAXS experiments using a sliding window of a user-specified size, calculates overall parameters [$I(0)$, R_g, D_{max} and molecular weight] and predicts the folding state (folded/unfolded) of the sample. Applications of *DATASW* are illustrated for several proteins with various oligomerization behaviours recorded on different beamlines. *DATASW* binaries for major operating systems can be downloaded from http://datasw.sourceforge.net/.

1. Introduction

The application of small-angle X-ray scattering (SAXS) to study the structure and properties of biological macromolecules in solution has seen exponential growth over the last decade (Graewert & Svergun, 2013). SAXS can provide a rapid assessment of parameters such as molecular weight (MW), radius of gyration (R_g) and the overall shape of a macromolecule, thereby complementing high-resolution crystallography or nuclear magnetic resonance studies (Elegheert *et al.*, 2012; Lapinaite *et al.*, 2013). SAXS can be particularly useful for the characterization of oligomerization-prone (Soderberg *et al.*, 2013) and flexible (Shkumatov *et al.*, 2011) proteins.

If a solution contains multiple macromolecule species, the measured SAXS curve will be an average of all these species. Deconvolution of such curves is far from trivial (Petoukhov *et al.*, 2012). A more efficient separation can be achieved by passing the sample through an HPLC system while the SAXS signal from the eluate is continuously recorded (Mathew *et al.*, 2004). The existing tools for the analysis of HPLC–SAXS data include automatic data-analysis pipelines that are typically run in real time during data collection, allowing very limited adjustment by the user, and standalone programs such as *FOXTROT* (David & Pérez, 2009) and *US-SOMO* (Brookes *et al.*, 2010). *FOXTROT*, developed at the SWING beamline, can only perform frame-per-frame buffer subtraction and calculation of R_g and $I(0)$ using fixed data range for all frames specified by the user. *US-SOMO* is a sophisticated analysis and modelling framework for various experimental data recently supplemented by an HPLC–SAXS module, albeit still under development (Brookes *et al.*, 2013). Overall, *US-SOMO* can be quite cumbersome for novice SAXS users.

DATASW was developed for the rapid analysis of subtracted HPLC–SAXS, calculating for each data frame the

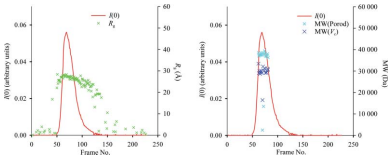

Table 1
DATASW strategy to process subtracted HPLC–SAXS data.

Action	Processing tool	Output file name
Frame averaging using user-specified sliding window	*DATAVER*	–
Automatic computation of R_g and $I(0)$ using the Guinier approximation, filtering out low-quality frames based on R_g quality estimate	*AUTORG*	`autorg_summary.txt`; `autorg_summary-Peak.txt`; `Frame-versus-I0Rg.txt`
Prediction of folding state (folded/unfolded)	*DATCLASS*	`autorg_summary-Peak.txt`
Peak detection based on $I(0)$ values	Built-in†	
Estimation of D_{max}, the distance distribution function $P(r)$ and the regularized scattering curve	*DATGNOM*	`MW_estimation.txt`; `Frame-versus-MWs.txt`
Computation of Porod volume from the regularized scattering curve and MW	*DATPOROD*	
Estimation of MW from the volume of correlation	*DATVC*	
Generation of plots [frame *versus* $I(0)/R_g$ and frame *versus* MW(Porod)/MW(V_c)]	*GNUPLOT*†	`frame-versus-i0rgmw.pdf` or `frame-versus-i0rgmw.ps` (Unix)

† With the exception of these two tools, the tools are from the *ATSAS* package (Petoukhov *et al.*, 2012).

zero-angle scattering $I(0)$, the radius of gyration R_g, the maximal dimension D_{max} and two estimates of the MW from the Porod volume V_p and the volume of correlation V_c (Rambo & Tainer, 2013), as well as suggesting the folding state (folded/unfolded). *DATASW* plots calculated parameters as a function of frame number and a publication-ready figure is generated.

2. Implementation and requirements

DATASW is written in Perl. The program supports various datafile-naming formats used at SAXS beamlines. *DATASW* requires tools from the *ATSAS* (v.2.5.2 or higher) program package (Petoukhov *et al.*, 2012; Table 1). For peak detection, a simple built-in algorithm is used: a baseline is calculated as an average over computed $I(0)$ values for all subtracted frames; thereafter, groups of frames with $I(0)$ values above the mean value are identified as peak areas. MW estimation and peak detection are only performed for frames where R_g can be reliably estimated by *AUTORG* (Table 1), *i.e.* the R_g quality estimate must be higher than 10% (where 0% corresponds to unusable data and 100% to data of ideal quality) with a standard deviation of no more than 30%. The calculated Porod volume is divided by 1.7 to obtain an estimate of the molecular weight (Petoukhov *et al.*, 2012). *GNUPLOT* (http://www.gnuplot.org) is used to generate a figure containing plots of R_g, $I(0)$ and MW *versus* frame number.

3. Features

An HPLC–SAXS experiment results in a set of files with intensity *versus* scattering angle recorded as a series of 'frames' with user-defined exposure time. Final frames need to be buffer-subtracted, as performed by automatic pipelines at the BM29 and P12 beamlines or manually by the user: the *FOXTROT* software is used at the SWING beamline. As an input for *DATASW*, one or more folders with subtracted datafiles are used. Typically, buffer-subtracted datafiles should

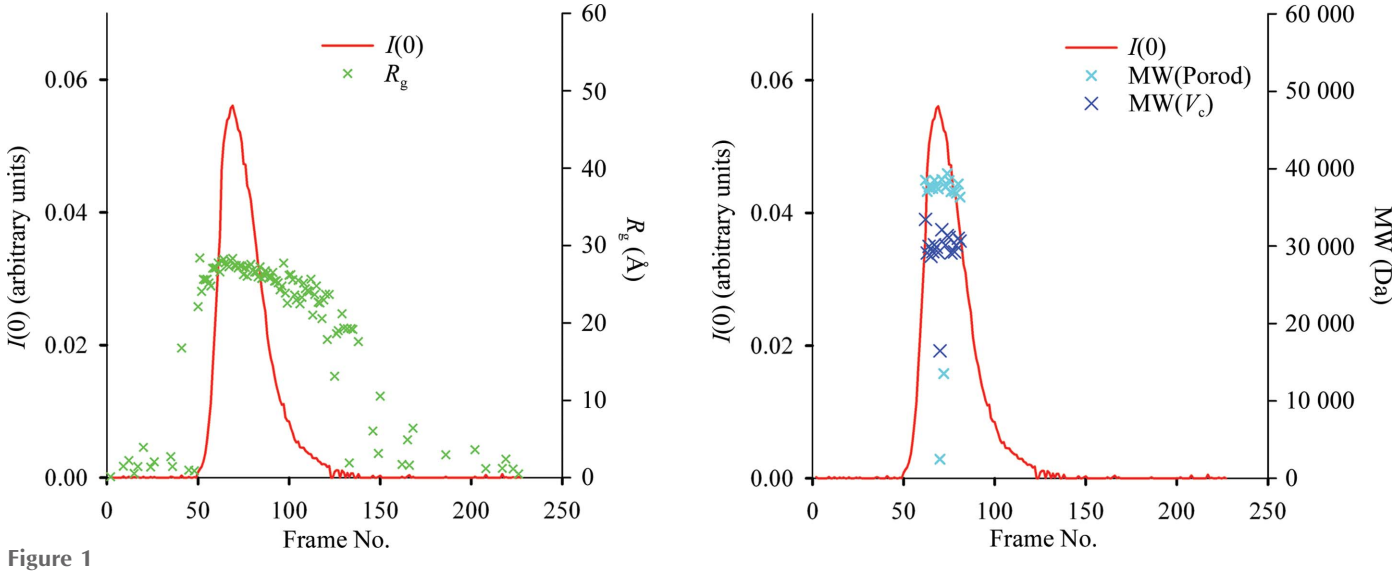

Figure 1
Output generated by *DATASW* for a typical monodisperse sample (HSPB6 dimer; Weeks *et al.*, 2014). The forward scattering $I(0)$ and radius of gyration R_g are shown as a red line and green crosses, respectively (left). MW determined from V_p (cyan crosses) and V_c (blue crosses) are plotted *versus* frame number (right). MW is only calculated for the frames of the automatically determined peak.

have a common identifier optionally followed by some text, an underscore separating a sequential number and a `.dat` extension. It is recommended that the files should be prepended with a common identifier, *e.g.* 'hplc', to enable the processing of multiple folders. *DATASW* can be run in interactive or batch mode, and for the latter the most important parameters can be supplied as arguments on the command line. The output includes a folder with subtracted data files, optionally averaged using a sliding window size specified by the user. This step may be necessary to increase the signal-to-noise ratio (SNR) in case the concentration of the injected sample was low. The output folder also contains files and figures reflecting calculated parameters (Fig. 1, Table 1). Peak detection can be disabled for oligomerization-prone samples to allow calculation of the MW and folding state for each sample frame where R_g can be estimated. Separate folder(s), named 'Peak_*', within the results directory contain frames corresponding to the detected peak(s). For each detected peak two files are created. One file contains an average over all frames in the peak and the second contains an average over ±10% of all frames in the peak with reference to the maximal $I(0)$ value.

4. Examples of *DATASW* usage

Here, we describe three examples of samples with different oligomerization behaviour showing: (i) a single peak profile, (ii) multiple peaks and (iii) a profile corresponding to progressive oligomerization. Sample preparation is described in the Supporting Information. The first example describes human small heat-shock protein B6 (HSPB6; Weeks *et al.*, 2014; Heirbaut *et al.*, 2014). HPLC–SAXS data recorded on the SWING beamline show a single, almost symmetrical peak for HSPB6 dimers. Notably, the R_g has a slightly bell-shaped profile across the peak (Fig. 1), suggesting attractive interparticle interaction, as reported previously (Weeks *et al.*, 2014). For data averaging, we used a sliding window size of 3 to increase the SNR. The resulting SAXS curve was essentially identical to the curve reported previously (Weeks *et al.*, 2014). MW values calculated from the Porod volume and the volume of correlation are somewhat overestimated and underestimated, respectively (Fig. 1), although both are close to the expected MW of a dimer (~34 kDa).

The second example shows HPLC–SAXS data for a complex of the 14-3-3 protein with a binding partner (Sluchanko *et al.*, in preparation) recorded on the BM29 beamline. The high concentration of the loaded sample

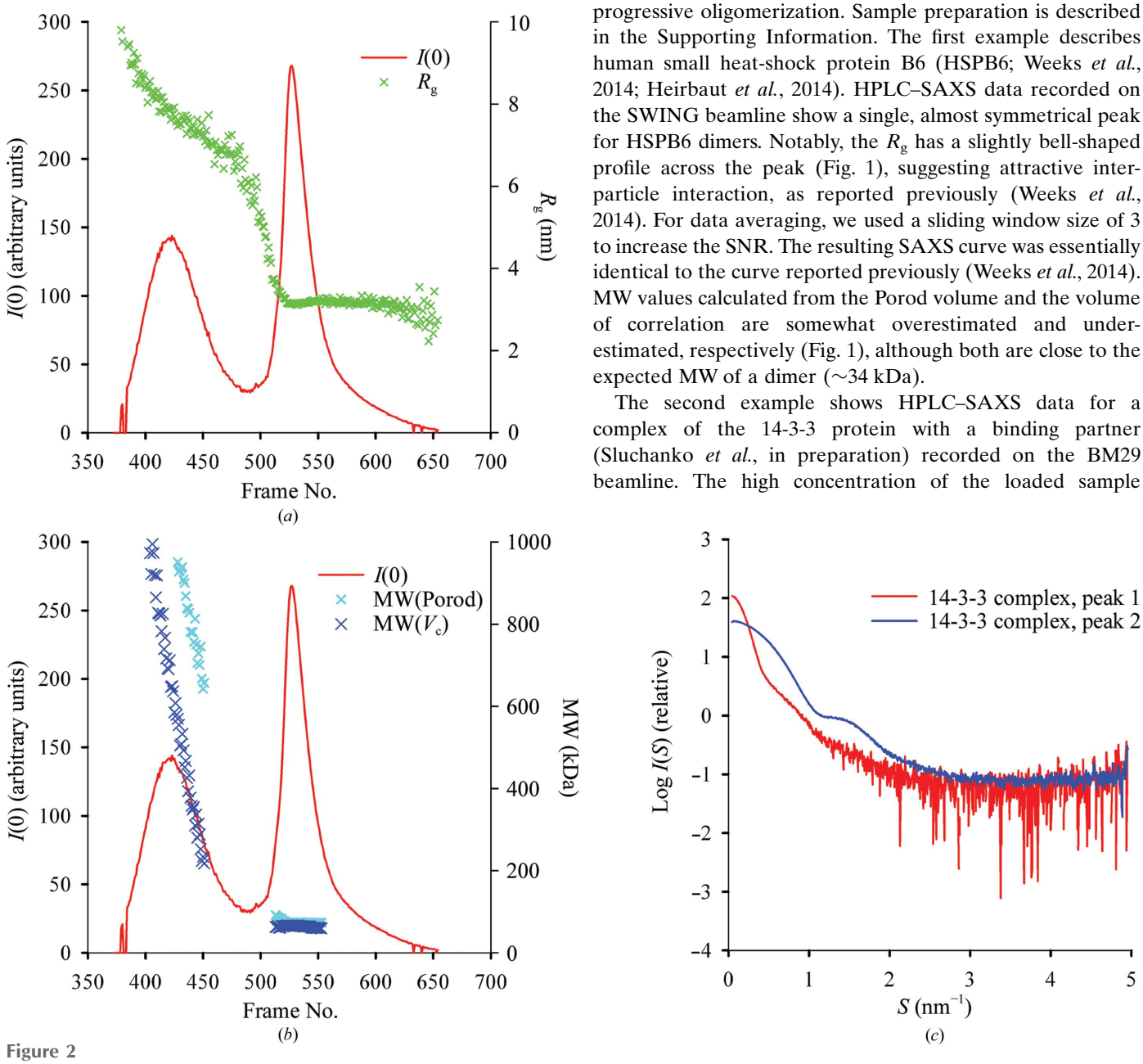

Figure 2
(*a*, *b*) *DATASW* output for a sample (14-3-3 protein complex) containing a high-molecular-weight oligomer (peak 1) alongside the main species (peak 2). (*c*) SAXS profiles for peak 1 (red line) and peak 2 (blue line).

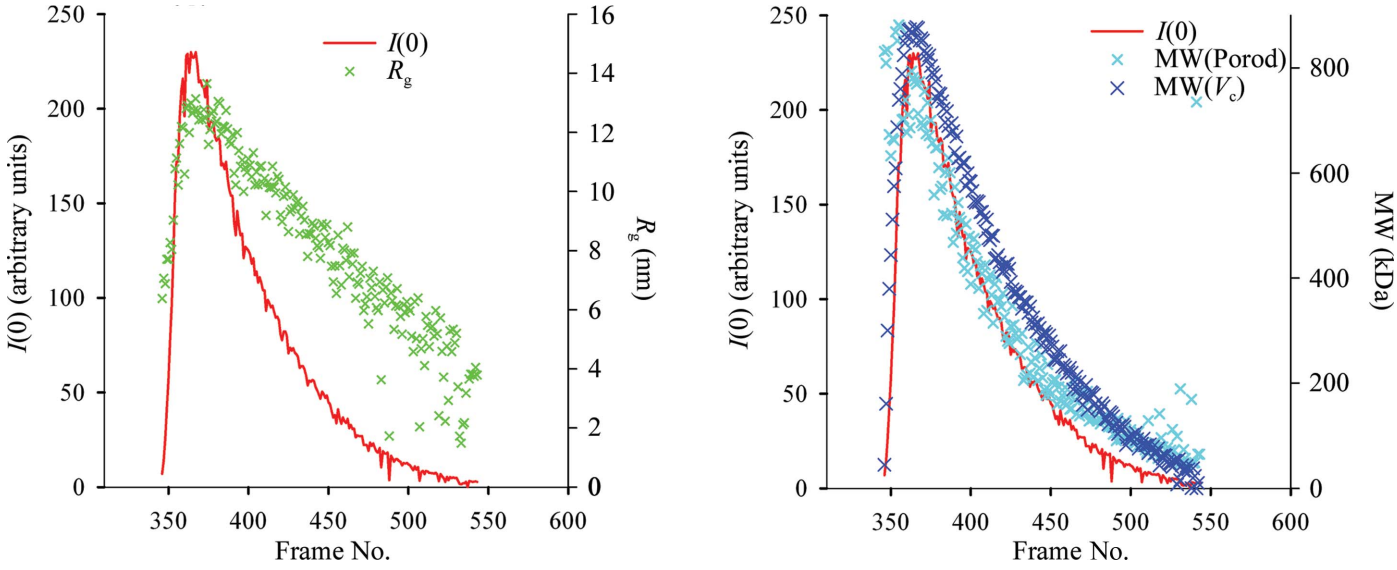

Figure 3
DATASW output for a sample (HSPB5 truncation) exhibiting progressive oligomerization. Peak detection was switched off to allow MW calculations for each sample frame.

(~12 mg ml^{-1}) and the use of the PILATUS 1M detector resulted in frames with a good SNR, hence additional frame averaging was not performed (sliding window size = 1). The expected MW of the complex is 85 kDa. The first detected peak appears to correspond to a polydisperse oligomer with a high MW (from ~1700 kDa down to ~650 kDa), indicated by a steep decrease in both R_g and MW with frame number (Fig. 2). The second detected peak apparently corresponds to a single species with a stable R_g value and an estimated MW of 73.5 ± 1 and 64.3 ± 2 kDa from V_p and V_c, respectively (Fig. 2). A more accurate estimation of MW from the excluded volume (Supporting Information) suggests a MW of 86 ± 1 kDa, which is in excellent agreement with the expected value.

The last example is a deletion mutant of HSPB5 (Shkumatov *et al.*, in preparation). The SEC profile exhibited a single asymmetric peak with a long trailing right shoulder, suggesting progressive oligomerization of this sample. *DATASW* was run without data-frame averaging and with peak detection disabled to perform MW estimation for every sample frame (Fig. 3). MW derived from the Porod volume decreased from ~880 to 45 kDa (the MW of a monomer is ~11.3 kDa).

In summary, *DATASW* is a new user-friendly tool to analyse subtracted HPLC–SAXS data and produce publication-ready plots within minutes.

Acknowledgements

We acknowledge the European Synchrotron Radiation Facility and Synchrotron SOLEIL for the provision of beam time. We would like to thank Dr M. Brennich and Dr E. Poudevigne for their help at the BM29 SAXS beamline. We are grateful to Dr A. Thureau for his help during measurements and data processing at the SWING beamline. We thank Dr N. Sluchanko, Dr Y. Sterckx and D. Guzenko for feedback and useful comments. Dr R. Sethi, Dr S. Weeks and C. Laverty are acknowledged for proofreading the manuscript. This research was supported by the KU Leuven (F+ fellowship to AVS and research grant OT13/097 to SVS), by the Research Foundation Flanders (FWO) grant G.0936.15 to SVS and by the European Community's Seventh Framework Programme under the BioStruct-X project.

References

Brookes, E., Demeler, B., Rosano, C. & Rocco, M. (2010). *Eur. Biophys. J.* **39**, 423–435.
Brookes, E., Pérez, J., Cardinali, B., Profumo, A., Vachette, P. & Rocco, M. (2013). *J. Appl. Cryst.* **46**, 1823–1833.
David, G. & Pérez, J. (2009). *J. Appl. Cryst.* **42**, 892–900.
Elegheert, J., Bracke, N., Pouliot, P., Gutsche, I., Shkumatov, A. V., Tarbouriech, N., Verstraete, K., Bekaert, A., Burmeister, W. P., Svergun, D. I., Lambrecht, B. N., Vergauwen, B. & Savvides, S. N. (2012). *Nature Struct. Mol. Biol.* **19**, 938–947.
Graewert, M. A. & Svergun, D. I. (2013). *Curr. Opin. Struct. Biol.* **23**, 748–754.
Heirbaut, M., Beelen, S., Strelkov, S. V. & Weeks, S. D. (2014). *PLoS One*, **9**, e105892.
Lapinaite, A., Simon, B., Skjaerven, L., Rakwalska-Bange, M., Gabel, F. & Carlomagno, T. (2013). *Nature (London)*, **502**, 519–523.
Mathew, E., Mirza, A. & Menhart, N. (2004). *J. Synchrotron Rad.* **11**, 314–318.
Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D. T., Konarev, P. V. & Svergun, D. I. (2012). *J. Appl. Cryst.* **45**, 342–350.
Rambo, R. P. & Tainer, J. A. (2013). *Nature (London)*, **496**, 477–481.
Shkumatov, A. V., Chinnathambi, S., Mandelkow, E. & Svergun, D. I. (2011). *Proteins*, **79**, 2122–2131.
Soderberg, C. A., Rajan, S., Shkumatov, A. V., Gakh, O., Schaefer, S., Ahlgren, E. C., Svergun, D. I., Isaya, G. & Al-Karadaghi, S. (2013). *J. Biol. Chem.* **288**, 8156–8167.
Weeks, S. D., Baranova, E. V., Heirbaut, M., Beelen, S., Shkumatov, A. V., Gusev, N. B. & Strelkov, S. V. (2014). *J. Struct. Biol.* **185**, 342–354.