Structure analysis of macromolecular solutions with small-angle X-ray scattering

Dmitri Svergun

EMBL
Biological SAXS @ EMBL-HH

Group leader: D. Svergun


Postdocs: A. Tuukkanen, M. Graewert, A. Spilotros, A. Panjkovich, M. Schroer, A. Gruzinov

Predocs: N. Hajizadeh, K. Malanastas

Major tasks:

- Development of data analysis methods
- Running and developing SAXS beamlines
- User support and collaborative projects
- Interactions, education and training
Small-angle scattering: experiment

Radiation sources:
- X-ray generator ($\lambda = 0.1 - 0.2 \text{ nm}$)
- Synchrotron ($\lambda = 0.03 - 0.35 \text{ nm}$)
- Thermal neutrons ($\lambda = 0.2 - 1 \text{ nm}$)

Monochromatic beam

Wave vector $k$, $k = 2\pi/\lambda$

Scattering vector $s = k_1 - k$
$s = 4\pi \sin \theta / \lambda$

Detector
Sample

$\theta$

Log (Intensity)

$0$ $1$ $2$ $3$

0

$2\theta$

$k_1$

$s = 4\pi \sin \theta / \lambda, \text{ nm}^{-1}$
Small-angle scattering: solvent

To obtain scattering from the particles, matrix scattering must be subtracted, which also permits to significantly reduce contribution from parasitic background (slits, sample holder etc).

Contrast $\Delta \rho = \langle \rho(r) - \rho_s \rangle$, where $\rho_s$ is the scattering density of the matrix, may be very small for biological samples.
Crystal solution
For SAXS solution studies, one does not need to grow crystals.

SAXS is not limited by molecular mass and is applicable under nearly physiological conditions.

Using solution SAXS, one can more easily observe responses to changes in conditions.

SAXS permits for quantitative analysis of complex systems and processes.

In solution, no crystallographic packing forces are present.
Small-angle scattering in structural biology

Radiation sources:
- X-ray tube (\(\lambda = 0.1 - 0.2 \text{ nm}\))
- Synchrotron (\(\lambda = 0.05 - 0.5 \text{ nm}\))
- Thermal neutrons (\(\lambda = 0.1 - 1 \text{ nm}\))

Data analysis
- Shape determination
- Rigid body modelling
- Missing fragments
- Oligomeric mixtures
- Hierarchical systems
- Flexible systems

Additional information:
- Scattering curve \(I(s)\)
- Resolution, nm:
  - 3.1
  - 1.6
  - 1.0
  - 0.8

SAXSMAN © A.Kikhney
Scattering from dilute macromolecular solutions (monodisperse systems)

The scattering is proportional to that of a single particle averaged over all orientations, which allows one to determine size, shape and internal structure of the particle at low (1-10 nm) resolution.
Overall parameters

Radius of gyration $R_g$ (Guinier, 1939)

$$I(s) \cong I(0) \exp\left(-\frac{1}{3} R_g^2 s^2\right)$$

Molecular mass (from $I(0)$)

Maximum size $D_{max}$: $p(r)=0$ for $r > D_{max}$

Excluded particle volume (Porod, 1952)

$$V = 2\pi^2 I(0)/Q; \quad Q = \int_0^\infty s^2 I(s) ds$$
Low resolution structures of macromolecules in solution

Shape and conformational changes of macromolecules and complexes

Validation of high resolution models and oligomeric organization

Rigid body models of complexes using high resolution structures

Addition of missing fragments to high resolution models
The scattering is related to the shape (or low resolution structure)

- Solid sphere
- Hollow sphere
- Flat disc
- Long rod
- Dumbbell
How to reconstruct 3D from 1D

DAMMIN: uses beads packed on a regular grid and simulated annealing to generate a (most possible) compact model fitting the experimental data
DAMMIF, a fast DAMMIN

DAMMIF is a completely reimplemented DAMMIN written in object-oriented code

- About 25-40 times faster than DAMMIN (in fast mode, takes about 1-2 min on a PC)
- Employs adaptive search volume
- Makes use of multiple CPUs

Ab initio dummy residues model

- Proteins typically consist of folded polypeptide chains composed of amino acid residues.

At a resolution of 0.5 nm a protein can be represented by an ensemble of $K$ dummy residues centered at the $C\alpha$ positions with coordinates $\{r_i\}$.

Scattering from such a model is computed using the Debye (1915) formula.

Starting from a random model, simulated annealing is employed similar to DAMMIN.
GASBOR run on C subunit of V-ATPase

Starting from a random “gas” of 401 dummy residues, fits the data by a locally chain-compatible model.
GASBOR run on C subunit of V-ATPase

Beads: Ambruster et al.
(2004, June)
FEBS Lett. 570, 119

Cα trace: Drory et al.
(2004, November),
EMBO reports, 5, 1148
Some words of caution

Or Always remember about ambiguity!
Shape determination of 5S RNA: a variety of DAMMIN models yielding identical fits

Program SUPCOMB – a tool to align and conquer

- Aligns heterogeneous high- and low-resolution models and provides a dissimilarity measure (NSD). For every point in the first model, the minimum value among the distances between this point and ALL points in the second model is found; the same is done for the second model. These distances are added and normalized against the average distances between the neighbouring points for the two models (computation time ~ N1*N2).

5S RNA: ten shapes superimposed

Most populated volume
Automated analysis of multiple models

1. Find a set of solutions starting from random initial seeds and superimpose all pairs of models with SUPCOMB.

2. Find the most probable model (which is on average least different from all the others) and align all the other models with this reference one.

3. Remap all models onto a common grid to obtain the solution spread region and compute the spatial occupancy density of the grid points.

4. Reduce the spread region by rejecting knots with lowest occupancy to find the most populated volume.

5. These steps are automatically done by a package called DAMAVER if you just put all multiple solutions in one directory.

5S RNA: final solution

The final model obtained within the solution spread region
When biologists go for SAXS

Care for a shape?

This is just trivial case: SAS yields much more
Modern life sciences widely employ hybrid methods

The most known and popular tool is, of course, Photoshop

SAXS also allows for a very effective hybrid model building where high resolution portions are positioned to fit the low resolution scattering data
Scattering from a macromolecule in solution

\[ I(s) = \left\langle |A(s)|^2 \right\rangle_\Omega = \left\langle |A_a(s) - \rho_s A_s(s) + \delta \rho_b A_b(s)|^2 \right\rangle_\Omega \]

- \( A_a(s) \): atomic scattering in vacuum
- \( A_s(s) \): scattering from the excluded volume
- \( A_b(s) \): scattering from the hydration shell


**CRYSON (neutrons):** Svergun et al. (1998) *P.N.A.S. USA*, **95**, 2267
The E2 cores of the dihydrolipoamide acyltransferase (E2) enzyme family form either octahedral (24-mer) or icosahedral (60-mer) assemblies. The E2 core from Thermoplasma acidophilum assembles into a unique 42-meric oblate spheroid. SAXS proves that this catalytically active 1.08 MDa unusually irregular protein shell does exists in this form in solution.

Principle of rigid body modelling

Using spherical harmonics, the amplitude(s) of arbitrarily rotated and displaced subunit(s) are analytically expressed via the initial amplitude and the six positional parameters: $C_{lm}(s) = C_{lm}(B_{lm}, \alpha, \beta, \gamma, x, y, z)$.

The scattering from the complex is then rapidly calculated as

$$I(s) = I_A(s) + I_B(s) + 4\pi^2 \sum_{l=0}^{\infty} \sum_{-l}^{l} \text{Re} \left[ A_{lm}(s) C_{lm}^*(s) \right]$$

A global refinement run with distance constraints

A tyrosine kinase MET (118 kDa) consisting of five domains

Program SASREF

Single curve fitting with distance constraints: C to N termini contacts

Addition of missing fragments

Moreover, addition of missing fragments can be combined with rigid body refinement (programs BUNCH and CORAL).

- Flexible loops or domains are often not resolved in high resolution models.
- Their tentative configuration can be reconstructed by fixing the known portion and adding the missing parts to fit the scattering from the full-length macromolecule.

Building native-like folds of missing fragments

- Using DR-type models and protein-specific penalty functions

**Primary sequence**  
**Secondary structure**  
**Excluded volume**

**Neighbors distribution**  
**Knowledge-based potentials**  
**Bond angles & dihedrals distribution**

Dynamics and function of the C-terminus of the *E. coli* RNA chaperone Hfq

The hexameric Hfq (HfqEc) is involved in riboregulation of target mRNAs by small trans-encoded RNAs. Hfq proteins of different bacteria comprise an evolutionarily conserved core, whereas the C-terminus is variable in length.

By bioinfomatics, NMR, synchrotron CD and SAXS the C-termini are demonstrated to be flexible and to extend laterally away from the hexameric core. The flexible C-terminal moiety is capable of tethering long and structurally diverse RNA molecules.

Addition of missing fragments: CORAL

- A merger of SASREF and BUNCH: advanced methods to account for missing loops in multi-subunit protein structures (RANLOGS, CORAL)

C-terminal domain of WbdD as a molecular ruler

In Escherichia coli O9a, a large extracellular carbohydrate with a narrow size distribution is polymerized from monosaccharides by a complex of two proteins, WbdA (polymerase) and WbdD (terminating protein).

A truncated construct WbdD\textsuperscript{1-459} is monomeric. For the construct WbdD\textsuperscript{1-556} MX yields an active trimer but AAs 505-556 are not seen in the crystal.

SAXS \textit{ab initio} shape reveals that the C-terminal is further extended. A rigid body model was constructed using coiled-coil C-terminal and refining the position of the catalytic domains.

\textit{In vivo} analysis of insertions and deletions in the coiled-coil region revealed that polymer size is controlled by varying the length of the coiled-coil domain.

Scattering from mixtures

\[ I(s) = \sum_k v_k I_k(s) \]

The scattering is proportional to that of a single particle averaged over all orientations, which allows one to determine size, shape and internal structure of the particle at low (1-10 nm) resolution. For equilibrium and non-equilibrium mixtures, solution scattering permits to determine the number of components and, given their scattering intensities \( I_k(s) \), also the volume fractions.
Flexible systems, interactions, mixtures and processes

Equilibrium oligomeric mixtures

Stoichiometry and complex formation

Natively unfolded proteins and multidomain proteins with flexible linkers

Protein folding/unfolding kinetics

Assembly/disassembly processes
Crystal structures of substrate-bound chitinase from *Moritella marina* and its structure in solution

Chitinases break down glycosidic bonds in chitin and only few crystal structures are reported because of the flexibility of these enzymes.

The dimeric crystal structure (at BESSY) of chitinase 60 from M. marina (MmChi60) contains four domains: catalytic, two Ig-like, and chitin-binding (ChBD). SAXS (at EMBL) demonstrates that MmChi60 is monomeric and flexible in solution. The flexibly hinged Ig-like domains may thus allow the catalytic domain to probe the surface of chitin.

Quantitative assessment of flexibility

- Automated classification (folded, partially or completely unfolded) is available
  
  D. Franke

- In ensemble methods, one generates a large pool covering the conformational space and selects sub-ensemble(s) fitting the available experimental data

- EOM 2.0 (G. Tria, 2015): advanced pool generation, e.g. use of (partial) point symmetry

- Quantification of flexibility using entropy and variation

EOM, Bernadó et al. (2007)

 ATSAS (All That SAS) roadmap

- World most comprehensive program suite for small-angle scattering data analysis from biomacromolecular solutions
- Consists of more than 25 programs developed at EMBL-HH since 1991
- Available for download since 1999, accessible online since 2006
- Presently, ATSAS has over 12,000 users from over 50 countries
- Online usage: 750 users/30,000 jobs per year
Recent ATSAS methods developments

**Correlation Map: quantification of data fitting and an alternative to \( \chi^2 \) when the experimental error estimates are not available**

**Automated determination of the useful data range by finding the number of reliable Shannon channels**

**Advanced ensemble analysis of flexibility: EOM version 2.0 including symmetry and quantitative characterization of the results**

**Intrinsic ambiguity of SAXS data: calculation of a propensity that a given scattering pattern yields ambiguous shape reconstruction**
Correlation map (CM)

Lysozyme shape determination shown as signs of residuals

With this distribution, CM provides a p-value saying whether the fit is statistically acceptable. Numerous simulations demonstrated that CM has essentially the same statistical power as $\chi^2$, but without the need of knowing the associated errors in the data.

D.Franke, C.Jeffries
Variability and resolution of \textit{ab initio} models

Analysis of the Fourier shell correlation functions between the multiple aligned models allows one (like in cryo-EM) to assess the resolution of the ensemble by identifying the point where the averaged correlation drops below 0.5. The approach is tested and validated in hundreds of model and real examples.

Program SASRES, available online and for download in ATSAS 2.8
Standardization, databases, web servers

Report of the wwPDB Small-Angle Scattering Task Force


DANESSA, an expert system for automated interpretation of a SAXS experiment given the data and available a priori information http://www.embl-hamburg.de/biosaxs/atsas-online/danessa.php (M. Petoukhov, in preparation)
### Table 2. SAXS Data collection and derived parameters for CD27L.

<table>
<thead>
<tr>
<th>Data collection parameters</th>
<th>CD27L (wild-type)</th>
<th>CD27L (C238R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instrument</td>
<td>EMEL X33 beam line (DORIS-III, DESY, Hamburg)</td>
<td>EM8L P12 beam line (PETRA-III, DESY, Hamburg)</td>
</tr>
<tr>
<td>Beam geometry</td>
<td>2.0 ±0.6 mm²</td>
<td>0.2 ±0.12 mm²</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>1.54</td>
<td>1.24</td>
</tr>
<tr>
<td>s range (Å⁻¹)</td>
<td>0.01–0.6</td>
<td>0.01–0.46</td>
</tr>
<tr>
<td>Exposure time (s)</td>
<td>8 × 15</td>
<td>1 × (20 ± 0.05 s)</td>
</tr>
<tr>
<td>Concentration range (mg/mL)</td>
<td>0.9–4.0</td>
<td>1.0–8.5</td>
</tr>
<tr>
<td>Temperature (K)</td>
<td>283</td>
<td>283</td>
</tr>
</tbody>
</table>

**Structural parameters**

| $I/I_0$ (relative) [from $p(r)$] | 44 ± 2 | 3653 ± 14 |
| $R_g$ (Å) [from $p(r)$]         | 33 ± 1 | 43 ± 2 |
| $I/I_0$ (cm⁻¹) [from Guinier]   | 45.6 ± 0.5 | 3664 ± 14 |
| $R_g$ (Å) [from Guinier]        | 33 ± 1 | 42 ± 1 |
| $D_{max}$ (Å)                   | 106    | 147     |
| Porod volume estimate (Å³)      | 72151 ± 1000 | 91690 ± 1000 |
| Excluded volume estimate (Å³)   | 94000 ± 10000 | 123000 ± 10000 |
| Dry volume calculated from sequence (Å³) | 39121/78219 (mon/dim) |      |

**Molecular mass determination**

| $I/I_0$ (cm⁻¹) BSA (66,000 Da) | 71.4 ± 0.4 | 37.61 ± 10 |
| Molecular mass M₁ (Da) [from $I/I_0$] | 42150 ± 5000 | 63780 ± 5000 |
| Molecular mass M₂ (Da) [from Porod volume ($V_p/1.6$)] | 45094 ± 5000 | 57306 ± 5000 |
| Molecular mass M₃ (Da) [from excluded volume ($V_{ex}/2$)] | 47000 ± 5000 | 61500 ± 5000 |

**Software employed**

- Primary data reduction: RADAVER
- Data processing: PRIMUS/Qr
- Ab initio analysis: DAMMIF
- Validation and averaging: DAMAVER
- Rigid-body modeling: CORAL
- Equilibrium analysis: OLIGOMER
- Computation of model intensities: CRYSOL
- 3D graphics representations: PyMOL, UCSF Chimera

Abbreviations: $M_1$: molecular mass; $R_g$: radius of gyration; $D_{max}$: maximal particle dimension; $V_p$: Porod volume; $V_{ex}$: Particle excluded volume.

$^a$Momentum transfer $s|s|=4\pi\sin(θ)/\lambda$.

$^{b}$Values reported for merged data sets (wild-type: 0.9 & 4.0 mg.mL⁻¹; C238R: 1 & 4 mg.mL⁻¹).

doi:10.1371/journal.ppat.1004228.002
Database development and submission curation is done in Hamburg. Presently offers 249 data sets and 394 macromolecular models (world largest).
DARA, a rapidly searchable database of over 150,000 SAXS patterns from the entire PDB
A Brand New Method and Server

Deciphering conformational transitions of proteins by small angle X-ray scattering and normal mode analysis

High brilliance EMBL SAXS beamline P12

- About $10^{13}$ ph/sec in 200*120 mm$^2$
- Energy between 4 and 20 keV (3.0 to 0.6 Å)
- Divergence below 0.05*0.05 mrad$^2$
- Multilayer monochromator mode: over $5 \times 10^{14}$ ph/sec
- SASFLOW pipeline for on-line data processing and analysis
- Full automation, remote and mail-in access

- Robotic EMBL/ESRF sample changer
- Automated FPLC/HPLC in parallel with biophysical sample characterisation
Highlights of EMBL SAXS user publications

- Human Muscle α-Actinin
- Conformational switch in collybistin
  - Soykan et al. EMBO J (2014)
- KD/SH2 domains of Abl kinase
- kLANA/DNA complexes
  - Ponnusamy et al. NAR (2015)

- Surface protein SASG
- WbdD as a molecular ruler
  - Hagelueken et al. NSMB (2015)
- SaThiM from vitamin B1 synthetic pathway
- Folded RTX Domain of CyaA
What does SAS tell about biological macromolecules

- Nothing known: *ab initio* low resolution structure
- Incomplete high resolution structure known: probable configuration of missing portions
- Complete high resolution structure known: validation in solution and biologically active oligomers
- High resolution structure of domains/subunits known: quaternary structure by rigid body refinement
- Mixtures/assemblies: volume fractions of components
- Flexible systems: quantitative analysis of configurational ensembles
What does SAS tell about biological macromolecules

- Nothing known: *ab initio* low resolution structure
- Incomplete high resolution structure known: probable configuration
- Complete high resolution structure known: validation in solution and biologically active oligomers
- High resolution structure of domains/subunits known: quaternary structure by rigid body refinement
- Mixtures/assemblies: volume fractions of components
- Flexible systems: quantitative analysis of configurational ensembles

Not just high throughput: important present and future applications of SAXS are functional complexes and processes (flexible, dynamic, transient, evolving), where SAXS is among the few methods providing quantitative structural information.
Some words of caution

- Always check your samples BEFORE doing SAS!
- Use the other methods and NEVER trust them blindly!
- Always check integral parameters BEFORE 3D modelling!
iNEXT – Infrastructure for Structural Biology

infrastructure for NMR, EM & X-rays for Translational research
Acknowledgments


- **Former BioSAXS Group members** and especially M. Petoukhov, P. Konarev, E. Valentini, M. Kachala, G. Tria

- **EMBL-HH:** R. Meijers, C. Loew, S. Fiedler, M. Wilmanns

- **Other EMBL units:** E. Lemke, G. Kleywegt, F. Cipriani

- **External collaborators:** J. Trewhella (Sydney), P. Tompa (Brussels), J. Westbrook (RCSB Rutgers), V. Volkov (Moscow), C. Betzel (Hamburg), M. Roessle (Luebeck), numerous users of P12