Solution X-ray Scattering from Biological Macromolecules

Javier Pérez
Beamline SWING, Synchrotron SOLEIL, Saint-Aubin, France
General Outline

• Introduction

• SAXS basics

• Data Analysis

• A few experimental considerations

• Modelling

• Conclusions
INTRODUCTION
Principles of Small Angle X-ray Scattering in solution

SAXS provides structural information about macromolecules in solution

- **Limits**
  - spherically averaged information \(\rightarrow\) low resolution
  - non unicity of the solution
  - does not distinguish elements in a mixture

- **Advantages**
  - solution (no crystal) \(\rightarrow\) kinetics, titration, T°, P
  - relatively easy to carry experiments
  - can be checked against atomic models

**SAXS is at its best when complementary (structural) information is available**
Principles of Small Angle X-ray Scattering in solution

Structural information directly obtained from a scattering curve

- biophysical parameters: size and type of shape (globular, multidomains, unfolded, …)
- molecular weight, oligomerization state and volume

3D structural modeling

- possible low resolution molecular shape (ab initio methods)
- direct comparison with high resolution model
- possible model of (un)structured missing parts
- rigid body orientations within multidomain structures

⇒ Models « compatible with SAXS data »
NOT unique models, NO electronic density maps.
Structural information about macromolecules in solution

Nothing known (except the curve)

Known or supposed all-atom models

Structure of subunits available

Zones of supposed high flexibility

Shape determination

Model validation / elimination

Rigid body modeling of the complex

Selection within an Ensemble of Random Conformations

DAMMIN
DAMMIF
DENFERT

CRYSOL
FOX5

SASREF
BUNCH
CORAL
DADIMODO

EOM
MES
What may solution scattering yield?

- Shape
- Internal structure
- Atomic structure

Resolution, nm

\[-1\]

\[0, 5, 10, 15\]

\[\log I(s)\]

\[2.00, 1.00, 0.67, 0.50, 0.33\]

Slide from Dmitri Svergun, EMBL Hamburg
SAXS BASICS
Elastic Thompson scattering by an electron

- What scatter X-rays are the electrons

X-ray incident beam
Wavelength $\lambda = 2\pi / k$

$$\vec{E}(R, t) = \frac{r_0}{R} |E_{0y} \cos(2\theta) e^{i(kR - \omega t)}| E_{0z}$$

$$\vec{E}_0(x, t) = \begin{bmatrix} E_{0y} \\ E_{0z} \end{bmatrix} e^{i(kx - \omega t)}$$

$$r_0 = \frac{e^2}{4\pi \varepsilon_0 mc^2} = 2.818 \times 10^{-15} m$$

- $r_0$ is called the electron classical radius

}\[\text{Solution X-ray Scattering from Biological Macromolecules} \quad \text{5th School of Saxs Data Analysis, 2016 May 2nd – 6th, LNLS, Campinas\]
Scattering amplitude by a particle

Coherent scattering: summing up amplitudes

- «Number» of electrons in volume $d^3r$ : $d\rho = \rho_e(r)d^3r$

Wave 1: $E_1(R,t) = -\frac{r_0}{R}E_0(2\theta)e^{i(kR-\omega t)}$

Wave 2: $E_2(R,t) = -\frac{r_0}{R}E_0(2\theta)e^{i(kR-\omega t\cdot k_i \cdot r - k_d \cdot r)}$

Phase shift between waves 1 and 2:

$$\Delta \varphi = k_i \cdot \vec{r} - k_d \cdot \vec{r} = (k_i - k_d) \cdot \vec{r} = -\vec{q} \cdot \vec{r}$$

- The scattered wave is the sum of the waves scattered by the electrons of «all the volumes $d^3r$»

Particle scattering «length» (or Amplitude):

$$A(\vec{q}) = -r_0 \int_V \rho_e(\vec{r})e^{-i\vec{q} \cdot \vec{r}} d^3r$$

$\vec{q}$ = Momentum transfer

$$q = \|\vec{q}\| = \frac{4\pi \sin(\theta)}{\lambda}$$
Intense scattered by a sample – Auto-correlation function

Scattering amplitude \[ A(q) = -r_0 \int V \rho_e(r) e^{-i \vec{q} \cdot \vec{r}} d^3r \]

Scattering intensity per unit volume : \( I(Q) \), usual unit: cm\(^{-1}\).

\[ I(q) = \frac{1}{V} A.A^* (q) = \frac{r_0^2}{V} \int \int V V \rho_e(r_1) e^{-i \vec{q} \cdot \vec{r}_1} \rho_e(r_2) e^{i \vec{q} \cdot \vec{r}_2} d^3r_1 d^3r_2 \]

\[ I(q) = \frac{r_0^2}{V} \int \int V V \rho_e(r_1) \rho_e(r_2) e^{-i \vec{q} \cdot (\vec{r}_1 - \vec{r}_2)} d^3r_1 d^3r_2 \]

Auto-correlation function \( \gamma_e(r) \):

\[ \gamma_e(r) = \frac{1}{V} \int V \rho_e(r') \rho_e(r + r') d^3r' \]

\[ I(q) = r_0^2 \int V \gamma_e(r) e^{-i \vec{q} \cdot \vec{r}} d^3r \]

The scattered intensity is the Fourier Transform of the electronic density auto-correlation function.
**Particles in a matrix (or buffer)**

- A particle is described by the associated electron density distribution $\rho_p(r)$.
- In a matrix, what contributes to scattering is the *contrast* of electron density between the particle and the matrix $\Delta \rho(r) = \rho_p(r) - \rho_0$ that may be **very small** for biological samples.

\[
\begin{align*}
\rho &= 0.43 \\
\rho_0 &= 0.334
\end{align*}
\]

\[\overline{\rho} = \frac{\rho + \rho_0}{2} = 0.387\]
Particules inserted in a "matrix"

- Scattering amplitude

\[
f(\vec{q}) = -r_0 \int_{V_1} \Delta \rho(\vec{r}) e^{-i\vec{q} \cdot \vec{r}} d^3 \vec{r}, \vec{q} \neq 0
\]

\(\Delta \rho(\vec{r})\) is the contrast of electronic density and describes the scattering object

\(f(\vec{q})\) is the Scattering Amplitude of the ensemble of the particles

- Scattering intensity per unit volume

\[
I(\vec{q}) = \frac{1}{V} f(\vec{q}) f^*(\vec{q})
\]

\(I(\vec{q})\) is expressed in cm\(^{-1}\) and is directly related to the measured intensity
Particles in solution

Particles in solution have random orientation, both in time (thermal motion) and in space (no long range correlations). The sample as a whole is therefore isotropic. As a result, the scattering intensity only depends on the modulus of $\vec{Q}$, $Q = 4\pi \sin(\theta) / \lambda$.

Scattering from a single particle in solution, averaged over time:

$$I_1(q) = \langle f_1(q) f_1^*(q) \rangle \Omega$$

The form factor $P(Q)$ is the normalized signature in $q$–space of a particle in solution.

$$P(q) = \frac{I_1(q)}{r_0^2 V_{obj} \left\langle \Delta \rho \right\rangle^2}$$

$$I_1(0) = r_0^2 V_{obj} \left\langle \Delta \rho \right\rangle^2$$

- Modulus
- Vector
- Average Electronic Density contrast
- Particle volume
Basic law of reciprocity in scattering

All, including large distances $\Delta r$ in the particle $\leftrightarrow$ Small scattering angle $q$

Short distances $\Delta r$ in the particle $\leftrightarrow$ Large scattering angle $q$

Phase: $q \cdot \Delta r$

$$I(q) = \frac{r_0^2}{V} \int \int_{V_1V_2} \Delta \rho_e(\vec{r}_1) \Delta \rho_e(\vec{r}_2) e^{-i\vec{q} \cdot (\vec{r}_1 - \vec{r}_2)} d^3r_1 d^3r_2$$
Basic law of reciprocity in scattering

Rotavirus VLP: diameter = 750 Å, 44 MDa

$D_{\text{max}} = 45$ Å

14.4 kDa

$I(Q)/c = \frac{4\pi \sin \theta}{\lambda} \left(\text{Å}^{-1}\right)$

Long distance correlations ↔ modulations at small q
A SAXS curve results from a pair of measurements: solution & buffer

\[ I_{\text{solution}}(q) - I_{\text{buffer}}(q) = I_{\text{particles}}(q) \]

To obtain scattering solely from the contrasting particles, intrinsic solvent scattering must be measured very accurately and subtracted, which also permits to subtract contribution from parasitic background (slits, sample holder etc) which should be reduced to a minimum.
Do not get mixed up!

Contrast effect

\[ \Delta \rho(\vec{r}) = \rho(\vec{r}) - \rho_0 \]

\[
 f(\vec{q}) = -r_0 \int_V \Delta \rho(\vec{r}) e^{-i\vec{q} \cdot \vec{r}} \, d^3r 
\]

\[ I(q) = \left\langle f(\vec{q}) f^*(\vec{q}) \right\rangle \]

Buffer subtraction

\[ I_{\text{particles,exp}}(q) = I_{\text{solution,exp}}(q) - I_{\text{buffer,exp}}(q) \]
Monodispersity and ideality

- **Monodispersity**
  - Yes ⇐ Identical particles
  - No ⇐ Size and Shape polydispersity

- **Ideality**
  - Yes ⇐ No correlations between particles positions
    (No short-range or long-range interactions)
  - No ⇐ Correlations between particles positions
    (Existence of short-range or long-range interactions)
Ideal and monodisperse solutions

\[ I(q) = \sum_{i=1,N} i_i(q) = \sum_{i=1,N} \left\langle f_i(\vec{q}) f_i^*(\vec{q}) \right\rangle_\Omega \]

\begin{itemize}
  \item Ideal
  \[ i_i(q) = i_1(q) \]

  \item Monodisperse
  \[ i_i(q) = i_1(q) \]

  \item Ideal and monodisperse
  \[ I(q) = N i_1(q) = N \left\langle f_1(\vec{q}) f_1^*(\vec{q}) \right\rangle_\Omega \]
\end{itemize}
One must check that both assumptions are valid for the sample under study.
Checking the validity of both assumptions for the sample under study is crucial for non erroneous data interpretation

- Size Monodispersity must be checked independently
  → Purification protocol: SEC, DLS, AUC, MALS, etc.

- Ideality: reached by working in buffers with screened interactions or at high dilution
  → In practice: measurements at decreasing concentrations and checks whether the scattering pattern is independent of concentration.
SE-HPLC / Solution Sampler

Flow rate 250 µl/min

- Monodisperse solution
- Aggregation is eliminated
- Oligomeric conformations can be distinguished
- Equilibrium states can be transiently separated
- Perfect background subtraction
- Automatic concentration series

Flow rate 5-40 µl/min

- Small volumes (~ 10 µl)
- No dilution
- High rate (a few minutes)

Incident X-ray

UV-Detector (280 nm)

Pump
Injection-mixing

Small molecules enter the aqueous spaces within beads

Large molecules cannot enter beads

Pure sample
DATA ANALYSIS
Data Analysis

• Guinier Analysis

• Kratky plot: why is it so interesting?

• « Real-space SAXS »: Distance correlation function \( P(r) \)
Data Analysis

• Guinier Analysis

• Kratky plot: why is it so interesting?

• «Real-space SAXS»: Distance correlation function $P(r)$
Data Analysis : Guinier law

Close to \( q=0 \), the scattering intensity of a particle can be described by a Gaussian curve.

The validity domain actually depends on the shape of the particle and is around \( q < 1.3 / R_g \) for a globular shape.

\[
I(q) = I(0) \exp \left( -\frac{q^2 R_g^2}{3} \right)
\]

Extrapolated intensity at origin

Radius of gyration

Guinier law, in Log scale:

\[
\ln[I(q)] = \ln[I(0)] - \frac{q^2 R_g^2}{3}
\]

The Guinier law is equivalent of a linear variation of \( \ln(I(q)) \) vs \( q^2 \) (Guinier plot). Linear regression on the experimental Guinier plot directly provides \( R_g \) and \( I(0) \).
**Guinier analysis**

\[ I(q) \approx I(0) - \frac{R_g^2}{3} q^2 \]

Validity range:

- \( 0 < R_g q < 1 \) for a solid sphere
- \( 0 < R_g q < 1.3 \) rule of thumb for a globular protein

Ideal monodisperse

\[ R_g = 27.8 \text{ Å} \]

\[ q R_g = 1.2 \]

\( q^2 \text{ (Å}^{-2}) \) vs. \( I(q) \)
**Radius of gyration**

\[ R_{g_{\text{exp}}}^2 = \frac{\int r^2 \Delta \rho(r) dr}{\int \Delta \rho(r) dr} \]

**Useful definitions of \( R_g \)**

\[ R_g = \frac{1}{N} \left\| \bar{r}_i - \bar{r}_{\text{COM}} \right\|^2 \]  
by atoms

\[ R_g^2 = \frac{\int \rho(r) r^2 dr}{\int \rho(r) dr} \]  
by electron density

\[ R_g = \frac{1}{2N(N-1)} \sum_{i} \sum_{j} \left| \bar{r}_i - \bar{r}_j \right|^2 \]  
by atom pairs

\[ R_g^2 = \frac{1}{2} \int r^2 \rho(r) dr / \int \rho(r) dr \]  
by pair distribution

**Lysozyme**

- \( R_g \) radius of gyration
- \( R_H \) hydrodynamic radius (not always \( > R_g \!\))
- \( R_R \) maximum hard sphere radius
- \( R_M \) radius of mass-equivalent sphere

* center of mass of the electron density

**Graph:** www.silver-colloids.com/Papers/hydrodynamic-radius.pdf

**Spheres:**
- \( R_g = \sqrt{\frac{3}{5}} R \)

**Thin rod:**
- \( R_g = \sqrt{\frac{1}{12}} L \)

**Thin disk:**
- \( R_g = \sqrt{\frac{1}{2}} R_{\text{disk}} \)
Mass retrieval from Guinier analysis

\[ I(Q) = I(0) \exp \left( -\frac{Q^2 R_g^2}{3} \right) \]

Absolute Unit: \( \text{cm}^{-1} \)

Classical electron radius

\[ I(0) = c \cdot \frac{M \cdot r_0^2}{N_A} \cdot \left[ v_p \left( \rho_{\text{prot}} - \rho_{\text{buf}} \right) \right]^2 \]

Mass concentration \( c \)

Electronic density contrast \( \rho_{\text{prot}} - \rho_{\text{buf}} \)

Protein specific volume \( v_p \)

\[ R_g^2 = \frac{\int_v r^2 \Delta \rho_{\text{prot}}(\vec{r}) d\vec{r}}{\int_v \Delta \rho_{\text{prot}}(\vec{r}) d\vec{r}} \]

Rg depends on the volume AND on the shape of the particle

I(0) gives an independent estimation of the molar mass of the protein (only if the mass concentration, \( c \), is precisely known ...)

Typically:

\[ M (\text{kDa}) = 1500 \times I_0 \text{ (cm}^{-1}) / C \text{ (mg/ml)} \]

For globular proteins: \( R_g (\text{Å}) \approx 6.5 \times M^{\frac{1}{3}} \), \( M \) in kDa

For unfolded proteins: \( R_g (\text{Å}) \approx 8.05 \times M^{0.522} \)

Bernado et al. (2009), Biophys. J., 97 (10), 2839-2845.
Example of Mass retrieval from Guinier analysis

Hen egg-white lysozyme

- C = 5.6 g/l
- Average of 8 frames of 2s
- Buffer subtracted
- Normalized by solid angle
- Normalized by transmitted intensity

\[ \ln[I(q)] = \ln[I(0)] - \frac{R_g^2}{3} q^2 \]

\[ R_g = 15.1 \pm 0.03 \, \text{Å} \]
\[ I_{\text{exp}}(0) = 0.0543 \, \text{cm}^{-1} \]

From \( I(0) \) provided the set-up was calibrated to give \( I(Q) \) in absolute units (cm\(^{-1}\)):

\[ M_{\text{exp}}(\text{kDa}) = I_{\text{exp}}(0) \times 1500 / c, \]
\[ \Rightarrow M_{\text{exp}} = 14.6 \, \text{kDa} \]

From \( R_g \), supposing the protein is globular:

\[ M_{Rg}(\text{kDa}) = \left( \frac{R_g}{6.3} \right)^3 \]
\[ \Rightarrow M_{Rg} = 13.8 \, \text{kDa} \]
Evaluation of the solution properties

Irreversible aggregation

→ Useless data: the whole curve is affected

I(0): > 150 fold the expected value for the given MM

(Courtesy D. Durand, IBBMC, Orsay)
Evaluation of the solution properties

Weak aggregation → possible improvement

centrifugation, buffer change

Nanostar –PR65 protein

\[ R_g \approx 38 \text{ Å} \text{ – too high!!} \]

\[ R_g \approx 36 \text{ Å} \]

(Courtesy D. Durand, IBBMC, Orsay)
Evaluation of the solution properties

Guinier plot

A linear Guinier plot is a requirement, but it is NOT a sufficient condition ensuring ideality (nor monodispersity) of the sample.
Evaluation of the solution properties

Guinier plot

same $R_g$ at all three concentrations

No interactions.

$N.\ Leulliot\ et\ al.,\ JBC\ (2009),\ 284,\ 11992-99$

$qR_g = 1.3$
Guinier plot

c = 4
R_g = 49.3 Å

RNA molecule
L. Ponchon, C. Mérigoux et al.
Evaluation of the solution properties

Guinier plot

c = 3
R_g = 56.6 Å

c = 4
R_g = 49.3 Å

RNA molecule
L. Ponchon, C. Mérigoux et al.
Guinier plot

c = 2
$R_g = 59.9 \, \text{Å}$
c = 3
$R_g = 56.6 \, \text{Å}$
c = 4
$R_g = 49.3 \, \text{Å}$

RNA molecule
L. Ponchon, C. Mérigoux et al.
Evaluation of the solution properties

Guinier plot

c = 1
R_g = 60.8 Å

c = 2
R_g = 59.9 Å

c = 3
R_g = 56.6 Å

c = 4
R_g = 49.3 Å

RNA molecule
L. Ponchon, C. Mérigoux et al.
Data Analysis

• Guinier Analysis

• Kratky plot: why is it so interesting?

• «Real-space SAXS»: Distance correlation function $P(r)$
**Kratky Plot**

SAXS provides a sensitive means to *evaluate the degree of compactness* of a protein:

- To determine whether a protein is globular, extended or unfolded
- To monitor the folding or unfolding transition of a protein

This is most conveniently represented using the so-called Kratky plot:

**Folded particle**: *bell-shaped curve* (asymptotic behaviour $I(Q)\sim Q^{-4}$)

**Random polymer chain**: *plateau* at large $q$-values (asymptotic behaviour in $I(Q)\sim Q^{-2}$)

**Extended polymer chain**: *increase* at large $q$-values (asymptotic behaviour in $I(Q)\sim Q^{-1.5}$)

Kratky Plots of folded proteins

Folded proteins display a bell shape. Can we go further?
Dimensionless Kratky Plots of folded proteins

For globular structures, DLKPs fold into the same maximum

The relation $M_Rg(kDa) \approx (Rg / 6.5)^3$ only works for the globular structures, not the elongated

The maximum value on the dimensionless bell shape tells if the protein is globular.
The bell shape vanishes as folded domains disappear and flexibility increases.

The curve increases at large Q as the structure extends.
In practice, thin Gaussian chains do not exist.

In spite of the plateau at $T=76^\circ\text{C}$, NCS is not a Gaussian chain when unfolded, but a thick chain with persistence length.

Pérez et al., *J. Mol. Biol.* (2001), 308, 721-743
Molecular Weight estimation based on Porod invariant

http://www.ifsc.usp.br/~saxs/saxsmow.html

- does not require knowledge of concentration
- relies on Porod Volume theory + structural database
- does not work for proteins with unfolded domains

Recent methods for MW estimation based on similar though different grounds were developed
Data Analysis

• Guinier Analysis

• Kratky plot: why is it so interesting?

• «Real-space SAXS»: Distance correlation function $P(r)$
The distance distribution function $p(r)$ characterises the shape of the particle in real space. The distance distribution function $p(r)$ is proportional to the average number of atoms at a given distance, $r$, from any given atom within the macromolecule. 

$p(r)$ vanishes at $r = D_{\text{max}}$
Relation between \( p(r) \) and \( I(q) \)

Intensity is the Fourier Transform of self-correlation function \( \gamma_{\text{obj}}(r) \):

\[
I(q) = 4\pi r_e^2 \varphi \int \gamma_{\text{obj}}(r) r^2 \frac{\sin(qr)}{qr} dr
\]

And:

\[
p(r) = \gamma_{\text{obj}}(r) r^2
\]

Then:

\[
I(q) = 4\pi r_e^2 \varphi \int_0^D p(r) \frac{\sin(qr)}{qr} dr
\]

And:

\[
p(r) = \frac{r^2}{2\pi^2 \varphi r_e^2} \int_0^\infty q^2 I(q) \frac{\sin(qr)}{qr} dq
\]

\( p(r) \) could be directly derived from \( I(q) \). Both curves contain the same information.

However, direct calculation of \( p(r) \) from \( I(q) \) is made difficult and risky by \([Q_{\text{min}}, Q_{\text{max}}]\) truncation and data noise effects.
Main hypothesis: the particle has a « finite » size, characterised by $D_{\text{max}}$:

- $D_{\text{max}}$ is proposed by the user
- $p(r)$ is expressed over $[0, D_{\text{Max}}]$ by a linear combination of orthogonal functions

\[
p_{\text{theoret}}(r) = \sum_{1}^{M} c_n \varphi_n(r)
\]

- $I(q)$ is calculated by Fourier Transform of $p_{\text{theoret}}(r)$

\[
I(q) = 4\pi r_e^2 \varphi \int_{0}^{D_{\text{max}}} p_{\text{theoret}}(r) \frac{\sin(q \cdot r)}{q \cdot r} dr
\]

Svergun (1988): program "GNOM"

- $M \sim 30 - 100$ ⇒ ill-posed LSQ ⇒ regularisation method
- "Perceptual criteria": smoothness, stability, absence of systematic deviations

- Each criterium has a predefined weight
- The solution is given a score calculated by comparison with « ideal values »
Distance Distribution Function

Experimental examples

GBP1

Heat denaturation of Neocarzinostatin

Experimental examples

Topoisomerase VI

M. Graille et al., Structure (2008), 16, 360-370.
Distance Distribution Function

Scattering curves obtained on different complexes Spire-Actin and Actin alone

<table>
<thead>
<tr>
<th>Complexes</th>
<th>Radius of gyration</th>
<th>Maximum diameter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>75.5 Å</td>
<td>285 Å</td>
</tr>
<tr>
<td></td>
<td>55.5 Å</td>
<td>210 Å</td>
</tr>
<tr>
<td></td>
<td>38.9 Å</td>
<td>130 Å</td>
</tr>
<tr>
<td></td>
<td>25 Å</td>
<td>75 Å</td>
</tr>
<tr>
<td></td>
<td>23.1 Å</td>
<td>70 Å</td>
</tr>
</tbody>
</table>

Histogram of intramolecular distances and ab initio molecular envelopes determined using DAMMIF
The radius of gyration and the intensity at the origin can be derived from \( p(r) \) using the following expressions:

\[
R_g^2 = \frac{\int_0^{D_{\text{max}}} r^2 p(r) dr}{2\int_0^{D_{\text{max}}} p(r) dr}
\]

and

\[
I(0) = 4\pi r_e^2 \varphi \int_0^D p(r) dr
\]

This alternative estimate of \( R_g \) makes use of the whole scattering curve, and is less sensitive to interactions or to the presence of a small fraction of oligomers.

**Comparison of estimates from Guinier analysis and from \( P(r) \) is a useful cross-check.**
A FEW EXPERIMENTAL CONSIDERATIONS
Schematics of beamline SWING

Vacuum chamber housing X-rays detectors

X-Z motorized table

Curved mirrors (KB config)

2 x Si(111) DCM
4.5 – 17 keV

Control room

Experimental Hutch

Optics Hutch

Full flux
5.10^{12} \text{ ph/s @ 12 keV}

Beam size (FWHM)
400 (H) x 25-100 (V) \mu m^2

In-vacuum Undulator U20
\text{gmin} = 5.5 \text{ mm}
Set-up for BioSAXS at Beamline SWING

Details of the BioSAXS cell

Online UV-Vis

Sample circulation

RX

BioSAXS Vacuum chamber

Incident beam

Detector chamber

SEC-SAXS

Solution X-ray Scattering from Biological Macromolecules

5th School of Saxs Data Analysis, 2016 May 2nd – 6th, LNLS, Campinas
Transmission and buffer measurements are crucial

- Transmission
  - The experimental scattering intensity must be normalised by transmitted intensity.
  - Transmission intensity must be measured with high accuracy (~ 0.1 %).

- Buffer
  - Buffer and protein samples must be measured in the same cell for correct subtraction of parasitic background arising from slits and holder walls.
  - The buffer in the buffer sample must be identical to that of the protein sample (dialysis, SEC, …).

\[ I_{\text{particles}}(Q) = I_{\text{sample}}(Q) - I_{\text{buffer}}(Q) \]
Particles in solution

Relation between the number of measured photons $\Delta N_{ph}$ on a given pixel of the detector, making a solid angle $\Delta \Omega$, and the Scattering Intensity per unit volume:

$$I(q) = \frac{1}{V} \frac{d\sigma}{d\Omega} (q) = \frac{\Delta N_{ph}}{N_0} \frac{1}{T \cdot e} \frac{D^2}{PxSize^2}$$

- Irradiated volume
- Differential cross-section
- Number of detected scattering photons in a given pixel
- Scattering Intensity per unit Volume
- Distance sample-pixel
- Sample thickness
- Sample transmission
- Number of incident photons
Calibration of the set-up using water scattering

Liquid scattering (theory): \( I(Q) = \text{constant at small } Q = r_0^2 Z^2 \rho_A^2 \cdot kT \kappa_T \)

\[
I_{\text{H}_2\text{O, theory}} = 0.0163 \text{ cm}^{-1}
\]

Water is used as primary reference to get the absolute intensity scale

- Capillary diameter = 1.6 mm
- Average of 2 frames of 2s
- Empty capillary subtracted
- Normalized by solid angle
- Normalized by transmitted intensity

Example:

\[
\begin{align*}
I_{\text{H}_2\text{O, exp}} &= 0.042 \text{ Exp. Units} \\
I_{\text{H}_2\text{O, exp}} &= K_{\text{exp}} \cdot I_{\text{H}_2\text{O, theory}}
\end{align*}
\]

\( \rightarrow \) Here : \( K_{\text{exp}} = 2.56 \text{ Exp. Units / cm}^{-1} \)

For any sample in that capillary : \( I_{\text{theory}}(\text{cm}^{-1}) = \frac{I_{\text{exp}}}{K_{\text{exp}}} = \frac{I_{\text{exp}}}{2.56} \)
Protocol for data collection and treatment

Data collection

1st case: the solution is supposed to be monodisperse

- Test radiation damage (7µl) $\rightarrow$ determine frame irradiation time

- Data collection on concentration series (25 µl) $\rightarrow$ take account of long range interactions

- Data collection on concentration series (25 µl) $\rightarrow$ take account of long range interactions
Protocol for data collection and treatment

Data treatment

- Subtract buffer → all curves I(Q)/c must superimpose at high Q
- Determine $I_0$ and $R_g$ → check for mass (aggregation ?) and long range interaction effects

- If necessary, merge low c (low Q) and high c (high Q) curves
- Compute $p(r)$ → should gently vanish at $D_{\text{max}}$
PRIMUS: combining data

- Small angle data using the lowest concentration curve or an extrapolation to zero concentration from a series of dilute solutions (correction of interparticle effects).
- Larger angle data using the most concentrated solution.

Here, slight repulsive interactions alter the concentrated curve at small angles.
The common range should be as restricted as possible to avoid adding noise.

“scale” function
PRIMUS: final merged curve

“merge” function
2nd case: the solution is a slow equilibrium or an unwanted mixture

- Use on-line HPLC data collection (typ 50 µl)

ASNP elution profile, monitored by UV absorption at 280 nm

Comparison between HPLC-purified and Direct injection curves

Fitting the HPLC-purified experimental curve with the crystal structure

I(0) and Rg determined for each SAXS frame during elution

Foxtrot program © SOLEIL
Fit of elution profiles by a set of gaussian curves
Each gaussian peak corresponds to an eluting species.

Following deconvolution, the scattering pattern 
$I_j(q)$ of each species $j$ can be reconstructed.

At this stage

We have gone from $I(Q)$ to $P(r)$.
Now, we have to go from

\[ I(Q) \]

\[ P(r) \]
MODELLING
The 1D SAXS profile is the Fourier transform of the 3D structure. But the inverse problem cannot be solved analytically, i.e., no “inverse computation” can be used to yield 3D position coordinates from scattering data.

How to reconstruct the 3D structure from the 1D SAXS profile?

Bear in mind!

One 3D structure → One SAXS curve
But
One SAXS curve → Many 3D structures, all compatible with the same curve
Additional constraints are always needed
SAXS data analysis, available programs

1) Nothing known (except the curve)

   ![Graph](image1.png)

   Low resolution model

   ![Graph](image2.png)

   DAMMIN
   DAMMIF
   GASBOR
   MONSA
   DENFERT

2) Theoretical model or complete atomic structure available

   ![Graph](image3.png)

   Validation/identification in solution

   ![Graph](image4.png)

   CRYSOLO
   FOXS

3) Structures of subunits available

   ![Graph](image5.png)

   Rigid body modeling of the complex and

   molecular modeling of the missing part

   ![Graph](image6.png)

   SASREF
   BUNCH
   CORAL
   DADIMODO
A word of caution

• \( s = 2 \sin \theta / \lambda \) modulus of the scattering vector

• \( Q = 2 \pi s = 4 \pi s \sin \theta / \lambda \) momentum transfer

• But in his programs: D. Svergun uses

\[ s = 4 \pi s \sin \theta / \lambda \]
Ab initio shape modelling: nothing is known but the curve
Ab initio shape modelling using a network of beads

Position\( (j) = X(j) = 1 \) or \( 0 \)

- \( M \approx \left( \frac{D_{\text{max}}}{r_0} \right)^3 \approx 10^3 >> N_s \) parameters, too many for conventional minimization
- No unique shape restoration unless constrained
- Able to describe complex shapes


• Obtaining 3D shapes from SAXS data is a ill-defined problem that can be **partially** solved by introducing additional information to **reduce** ambiguity of interpretation

• Using simulated annealing, finds a compact dummy atoms configuration $X$ that fits the scattering data by minimizing

$$f(X) = \chi^2[I_{\text{exp}}(s), I(s, X)] + \alpha P(X)$$

where $\chi$ is the discrepancy between the experimental and calculated curves, $P(X)$ is the penalty to ensure compactness and connectivity, $\alpha > 0$ its weight.

3D shape reconstructions from SAXS data with **DAMMIN**

compact

loose

disconnected
3D shape reconstructions from SAXS data with DAMMIN

- A series of runs (10-50) are performed to compare the different shapes obtained from the same data.

- After the run, an optimal superposition of models is realized with the program suite DAMSEL and DAMSUP.

- The algorithm defines a criteria of similarity, called « Normalized Spatial Discrepancy » or NSD, which measures the agreement between any pair of models.

- Similar shapes results in NSD < 1, very similar shapes NSD ≈ 0.5

- Models are conserved if its NSD < Mean of NSD + 2*standard deviation

- The model with the lowest NSD is the shape which has the most similarities with other, and **can** be regarded as the most representative of envelopes in accordance with the SAXS data

- Be careful with damfilt.pbd because $I_{\text{damfilt}}(q) \neq I_{\text{exp}}(q)$
Be aware: "Porod law" is forced for ab initio shape determination

DAMMIN: shape determination
Model with uniform density

Fitting data with approximate $q^{-4}$ high angle trend by subtracting a constant.
Ab initio model accounting for high resolution data

**DAMMIN/DAMMIF** : very low resolution, restricted portion of the data used ($q < 0.2 \text{ Å}^{-1}$), very basic constraints

**GASBOR** : a protein comprising $N$ residues is represented by an ensemble of $N$ spheres centered at the $C\alpha$ positions, the whole $q$-range can be used.

An initial gas-like distribution of dummy residues is refined using Simulated Anneling to fit the data under constraints ensuring a final chain like distribution.

*D. Svergun et al. (2001), Biophys. J., 80, 2946-2953.*
A word of caution: what NOT to do

- Common misconception: dummy atom ab initio envelopes are viewed as similar to EM density maps: **NO**.

- One should **not** try and superimpose 3D models of domains in the envelope. There is not 1 but MANY similar (or not) envelopes.

- One must try and refine the position of domains vs SAXS data.
From a atomic structure to a solution scattering pattern: program CRYSOL
Solvent scattering and contrast

$I_{\text{solution}}(q)$ $-$ $I_{\text{solvent}}(q)$ $=$ $I_{\text{particle}}(q)$

The bound solvent density differs from that of the bulk.

Bulk water density = 0.334 e-/Å³

Hydration layer density ~ 5-15 % higher
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, layer of thickness 3 Å


**CRYSON (neutrons):** Svergun et al. (1998) *P.N.A.S. USA*, 95, 2267
Program CRYSOL

• \( I(Q) \) is computed from the atomic coordinates.
• To gain computing time, \( I(Q) \) is developed in a series of Bessel functions and Spherical harmonics

\[
I_{\text{calc}}(Q) = \sum_{l=0}^{L} \sum_{m=-l}^{l} \left| A_{lm}(Q) - \rho_0 C_{lm}(Q) + \delta \rho B_{lm}(Q) \right|^2
\]

The experimental scattering curves are then fitted using only 3 parameters:
• the general scale of \( I_{\text{calc}}(Q) \)
• the total excluded volume \( V \), which is equivalent to modifying the average contrast.
• the contrast of the border layer \( \delta \rho \)

to minimize the discrepancy \( \chi \) :

\[
\chi^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left[ \frac{I_{\text{exp}}(Q_i) - \text{scale} \times I_{\text{calc}}(Q_i)}{\sigma_{\text{exp}}(Q_i)} \right]^2
\]

Effect of the hydration shell

T state of *E. coli* allostERIC ATCase

\[ Q = 4\pi \left( \frac{\sin \theta}{\lambda} \right), \text{Å}^{-1} \]

- Experimental data
- Fit without hydration shell
- Fit with hydration shell

\[ \gamma \]

\[ \alpha \]
Crysol application : Arf6
V. Biou et al., J.Mol.Biol (2010), 402(4), 696-707

Crystal structures of human Δ13Arf6-GDP (left) and Arf6-GDP-FullLength (right)

Δ13-Arf6 : a protruding loop in the crystal structure

Question : is the unfolded loop a crystal artifact ?
Crysol application: Arf6
V. Biou et al., J.Mol.Biol (2010), 402(4), 696-707

(a) Resolution (Å)

(b) Resolution (Å)

(c)

(d)

Solution X-ray Scattering from Biological Macromolecules
5th School of Saxs Data Analysis, 2016 May 2nd – 6th, LNLS, Campinas
From an atomic structure to a solution scattering pattern


*Most popular for BioSAXS, stand-alone program, fit model to data, fast computational algorithm. 1500 citations.*

http://www.embl-hamburg.de/biosaxs/atsas-online/crysol.php


*Use explicit water modeling solvation layer, robust fitting approach*

http://spin.niddk.nih.gov/bax/nmrserver/saxs1/


*A program to compute WAXS, Upon request*


*Debye-like computation, web server based. Hydration taken into account by “inflating” the volume of surface atoms.*

http://modbase.compbio.ucsf.edu/foxs/


http://waxis.uni-goettingen.de
When atomic structures of domains are known, but not their mutual arrangement
Rigid body modeling against SAXS data

SASREF: when atomic structures of domains are known, but no their mutual organization

The objective is to find the relative orientation and position of each subunit that gives a good agreement with the SAXS data of the complex.

The scattering intensity $I(q)$ of the complex is equal to the sum squared of the amplitudes of all subunits

$$ I(q) = \left\langle \left| \sum_{k=1}^{K} A^{(k)}(q) \right|^2 \right\rangle $$

$$ A^{(k)}(q) = \exp(iq \cdot \vec{r}_k) \prod (\alpha_{k} \cdot \beta_{k} \cdot \gamma_{k}) [C^{(k)}(q)] $$

Amplitudes are calculated with CRYSOL from the high resolution structure of each subunit.

The algorithm of minimization is the same used with DAMMIN with a penalty function (interconnectivity of the subunits, the steric clashes) and possibility to give information about contacting residues from other experiences.

$$ f(X) = \sum_i \chi_i^2 + \alpha_{dist} P_{dist}(X) + \beta_{cross} P_{cross}(X) + \gamma_{cont} P_{cont}(X) $$

Rigid body modeling with missing loop against SAXS data

**BUNCH and CORAL** : quaternary structure analysis of multidomain protein

**Combination of rigid body and ab initio modeling** :
- position and orientation of rigid domains
- possible conformation of flexible linkers

\[
f(X) = \sum_i \chi_i^2 + \alpha_{\text{ang}} P_{\text{ang}}(X) + \beta_{\text{cross}} P_{\text{cross}}(X) + \gamma_{\text{dih}} P_{\text{dih}}(X) + \delta_{\text{ext}} P_{\text{ext}}
\]

As SASREF, the amplitude are calculated with CRYSOL from the high resolution structure of each monomer

The algorithm of minimization is the same used with SASREF with a penalty function including the steric clashes Pcross, the dihedral angle Pang and Pdih, and the compactness of the loop Pext. The possibility to give information about contacting residues from other experiences is also added.

Flexibility → no unique structure!
NOT a structure but a SAXS data compatible model

ATSAS package and ATSAS online

http://www.embl-hamburg.de/biosaxs/software.html

http://www.embl-hamburg.de/biosaxs/atsas-online/

Solution X-ray Scattering from Biological Macromolecules

5th School of Saxs Data Analysis, 2016 May 2nd – 6th, LNLS, Campinas
DADIMODO: rigid body refinement vs. SAXS / NMR data

Collab: Christina Sizun & François Bontems (ICSN, Gif sur Yvette)

Modelling approach: complete atomic model
Full structure initiated with:
- Crystal or NMR domain structures
- Homology models

External information:
- Sequence
- Sub-parts moved as rigid-bodies (user-defined)
- A correct stereochemistry is maintained at all steps by minimizing energy (Amber 99 Force Field)

Experimental data:
- SAXS
- NMR
- RDC
- ADR (chem. shift map.)

Optimisation of the structure via a genetic algorithm
Dadimodo example: F45 from S1 protein

Structure:
- one polypeptide chain
- two rigid domains (D4 & D5) with known structures
- 1 linker
- 2 flexible parts (N-term and C-term)

Starting models population:
- obtained after running 50 steps Dadimodo with no selection pressure

Initial misfit to the SAXS data:
Dadimodo example: F45 from S1 protein

vs. SAXS only

vs. SAXS and ADR

vs. ADR only

vs. SAXS and RDC

vs. RDC only

vs. SAXS, ADR, RDC
CONCLUSIONS

• A scattering pattern can be calculated from atomic coordinates, thereby providing a link between crystal and solution work.

• Using SAXS patterns, *ab initio* methods can determine the shape of a molecule

• Rigid-body modeling allows one to propose models for complexes best fitting the data.

• Useful though limited structural information about flexible systems can be derived from SAXS data.
Comments

✓ SAXS is at his best when it is used to distinguish between several preconceived hypotheses.

✓ Analysis and modeling require a monodisperse and ideal solution, which has to be checked independently.

✓ Otherwise:

SAXS

IN

OUT
Elastic Thompson scattering by an electron

• What scatter X-rays are the electrons

\[ \vec{E}(r,t) = -\frac{r_0}{r} E_0 \begin{pmatrix} 0 \\ \cos(2\theta) \\ 1 \end{pmatrix} \cdot e^{i(kr-\omega t)} \]

X-ray incident beam
Wavelength \( \lambda = 2\pi / k \)
Flux \( I_0 = S \cdot E_0^2 \)

\[ \overrightarrow{E_0}(x,t) = E_0 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \cdot e^{i(kx-\omega t)} \]

\[ r_0 = \frac{e^2}{4\pi \varepsilon_0 mc^2} = 2.818 \times 10^{-15} m \]

\( r_0 \) is the electron classical radius

\( dI \): intensity scattered on the surface \( dS \) at distance \( r \)

\[ dI = dS \cdot E_0^2 \frac{r_0^2}{r^2} \left( \frac{1+\cos^2(2\theta)}{2} \right) = r^2 d\Omega \cdot E_0^2 \frac{r_0^2}{r^2} \left( \frac{1+\cos^2(2\theta)}{2} \right) \]

\[ b^2 = \frac{1}{E_0^2} \frac{dI}{d\Omega} = r_0^2 \left( \frac{1+\cos^2(2\theta)}{2} \right) \]

\( b_0 \) is the electron differential scattering cross section
SAXS experiments: strategy

**Guinier approximation**
- Rg (size) and I(0) (mass and oligomeric state)

**Distance distribution function p(r):**
- Dmax evaluation
- Rg (size) and I(0) compatibility with Guinier approximation
- Global form of the object

**Kratky plot**
- Type of structure (globular, elongated or unfolded)

**Porod law**
- Molecular volume if globular protein

**Cristallographic, NMR structures or complete molecular modeling**
- Theoretical curves calculation and data comparison

**Nothing is known**
- Low resolution shape

**Structures of subunits available**
- Molecular modeling rigid body against SAXS data

**Structures with missing loop or flexible parts**
- Molecular modeling of missing parts against SAXS data
From an atomic structure to a solution scattering pattern

\[ I_{th}(q) = \left[ A_a(q) - \rho_s A_s(q) + \delta \rho_b A_b(q) \right]^2 \]

\[ A_a(q) = \text{molecular scattering amplitude in vacuum} \]

\[ A_s(q) = \text{scattering amplitude from excluded volume} \]

\[ A_b(q) = \text{scattering amplitude from the hydration shell, layer of arbitrary thickness 3Å} \]

In **CRYSOL** program, in order to gain computing time, \( I(q) \) is developed in a series of Bessel functions and spherical harmonics:

\[ I_{\text{calc}}(q) = \sum_{l=0}^{L} \sum_{m=-l}^{l} \left| A_{lm}(q) - \rho_0 C_{lm}(q) + \delta \rho B_{lm}(q) \right|^2 \]

The experimental scattering curves are then fitted using only 3 parameters in order to minimize the discrepancy \( \chi^2 \):

- the general scale of \( I_{\text{calc}}(q) \)
- the total excluded volume \( V \), which is equivalent to modifying the average contrast \( \rho_0 \)
- the contrast of the border layer \( \delta \rho \)

\[ \chi^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left[ \frac{I_{\exp}(q_i) - \text{scale} \times I_{\text{calc}}(q_i)}{\sigma_{\exp}(q_i)} \right]^2 \]
3D shape reconstructions from SAXS data with DAMMIN

Ab initio shape modelling: nothing is known excepted the curve!

Principle of the method: any structure volume of homogeneous electronic density can be approximated at any resolution by a set of spheres of small enough diameter

Starting model = sphere with a radius \( R = D_{\text{max}}/2 \) with \( N \) scattered beads (\( r_0 \ll R \)

The number of the beads \( N \approx (R/r_0)^3 \)

Each bead is associated to a position \( j \) and an index \( X_j \) corresponding to the type of the phase (\( X_j = 0 \) for the solvent and \( X_j = 1 \) for the molecule)

\[
f(X) = \chi^2[I(q)_{\text{exp}}, I(q, X)] + \alpha P(X)
\]

\( X \) is a conformation of the system

\( P(X) \) is a penalty function

\[
\chi^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left[ \frac{I_{\text{exp}}(q_i) - \text{scale} \cdot I_{\text{calc}}(q_i)}{\sigma_{\text{exp}}(q_i)} \right]^2
\]

After \( k \) iterations