Solution X-ray Scattering from Biological Macromolecules

Javier Pérez

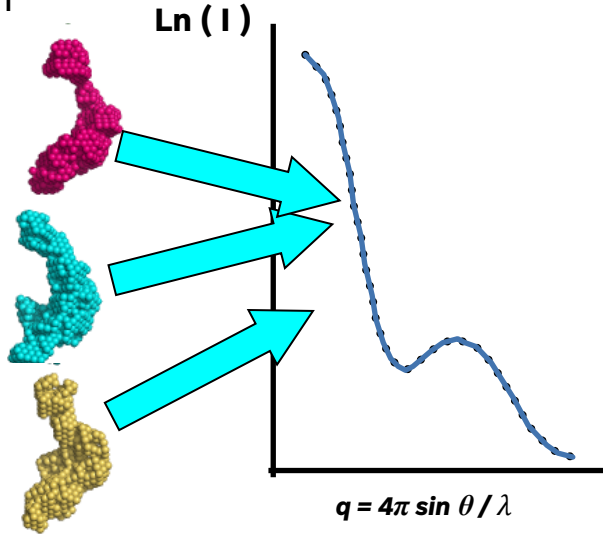
Beamline SWING, Synchrotron SOLEIL, Saint-Aubin, France

General Outline

- Introduction
- SAXS basics
- Biophysical information
- A few experimental considerations
- Modelling
- A few concluding messages

INTRODUCTION

Principles of Small Angle X-ray Scattering in solution



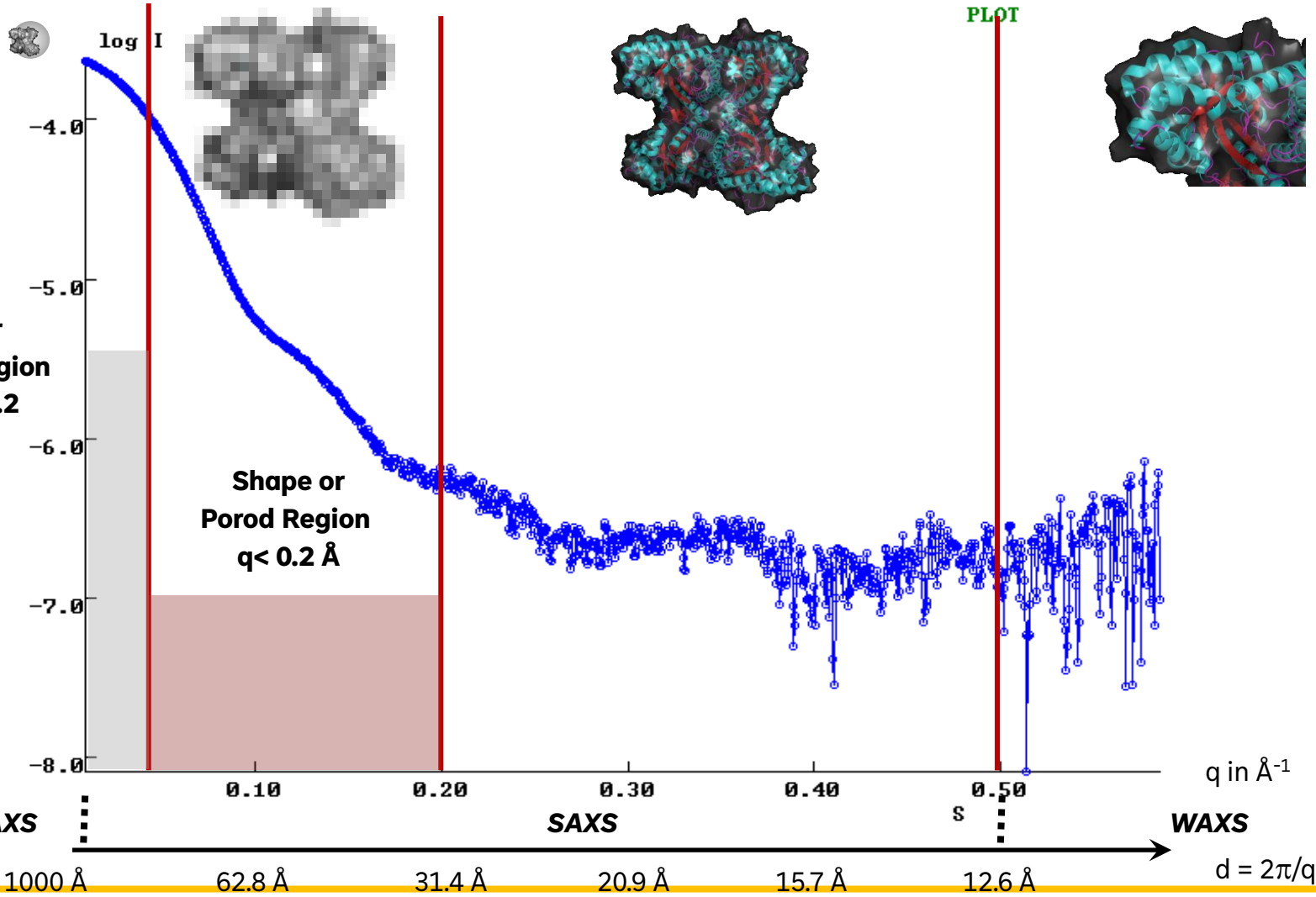
SAXS provides structural information about macromolecules in solution

- **Limits**
  - spherically averaged information → low resolution
  - non unicity of the solution
  - does not distinguish elements in a mixture
- **Advantages**
  - solution (no crystal) → kinetics, titration, T°, P
  - relatively easy to carry experiments
  - can be checked against atomic models

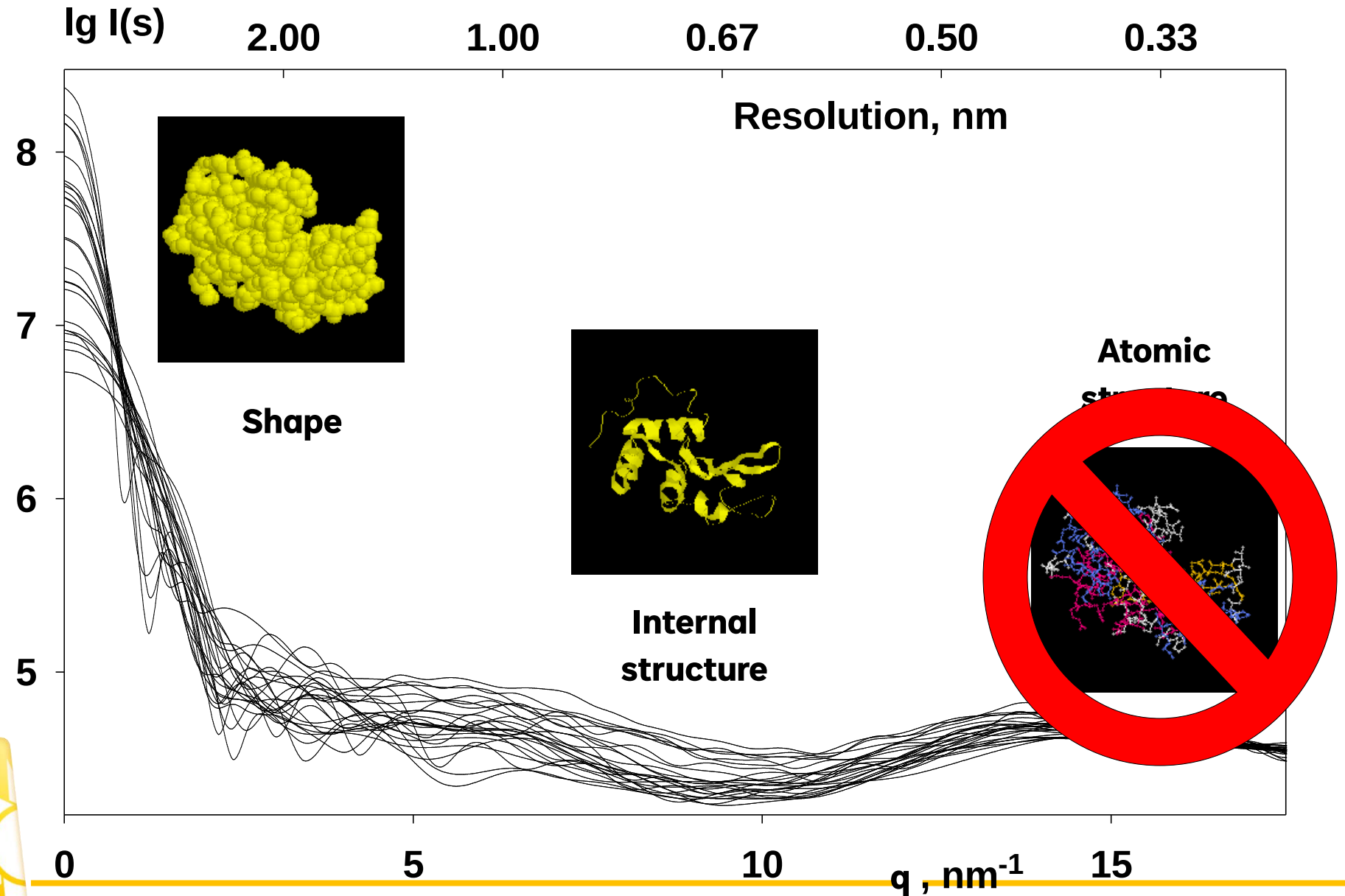
SAXS is at its best when complementary (structural) information is available

A typical BioSAXS curve from a protein

Global dimension → Global shape → Folding + domains → Secondary structures



What may solution scattering yield?



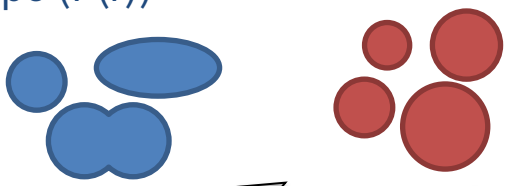
What may solution scattering yield?

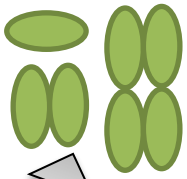
Shape (P(r))

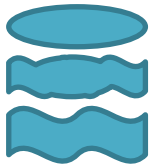
Size (Radius of gyration)

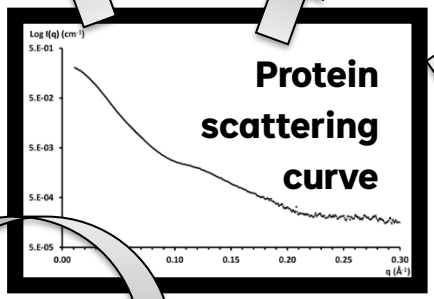
Mass / Oligomerisation
(~10% precision)

"Global flexibility"

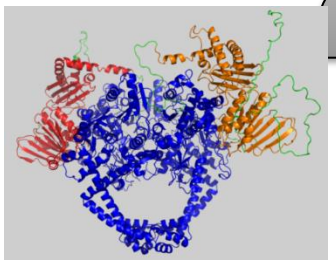




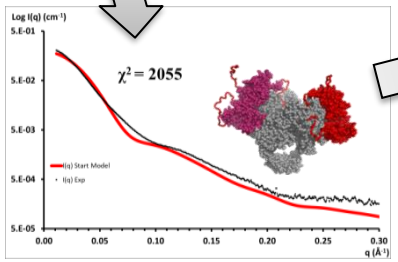




Atomic model from:
- MX
- NMR
- Cryo-EM
- Alphafold

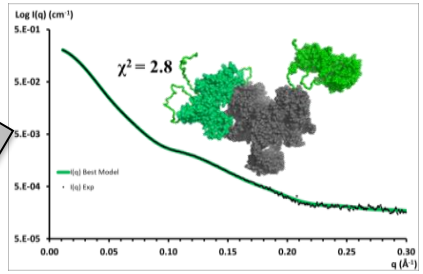


Petrella et al. 2019



Atomic Model comparison & validation

Molecular Modeling



Model "refinement"

Typical analysis steps

Direct Data analysis

Guinier fit

- Rg (size) & I(0) (mass and oligomeric state)

Pair distribution function p(r)

- Dmax evaluation
- Rg (size) and I(0) compatibility with Guinier approximation
- Global shape of the object

Kratky plot

- Type of structure (globular, elongated or unfolded)

Porod law

- Molecular mass if globular protein

Molecular modeling

Structural validation in solution

- *Complete atomic structures or AlphaFold models*

Ab initio modelling of compatible low resolution envelopes

- *No complementary structural information needed*

Inter-domain conformations

- *Atomic structures of subunits available*

Modeling of missing parts

- *Structures with missing loops or flexible parts*

→ « data compatible » models: NOT unique, NOT electronic density maps

SAXS BASICS

Elastic Thompson scattering by an electron

What scatter X-rays are the electrons

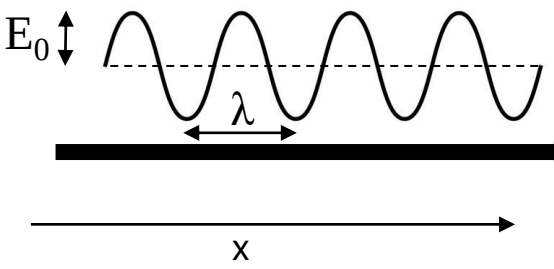
X-ray incident beam

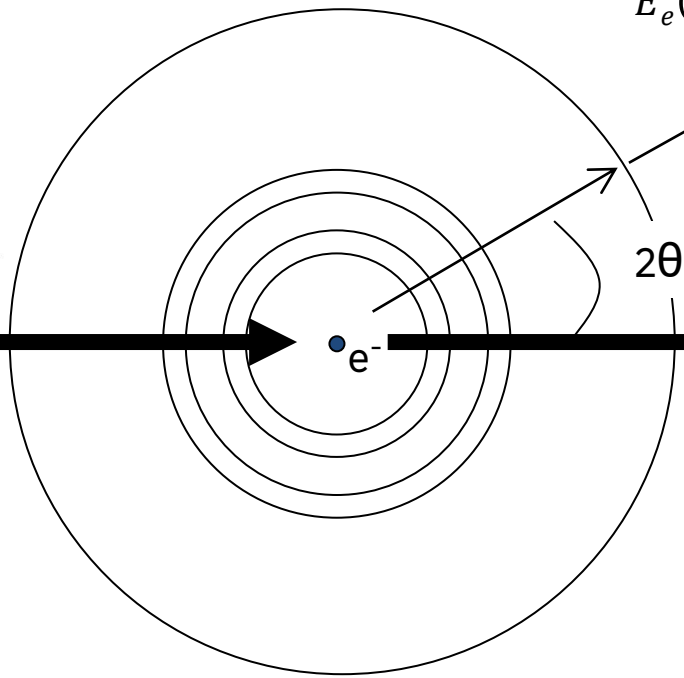
Wavelength : $\lambda \sim 1$ Å (10^{-10} m)

$k = 2\pi / \lambda$

$\omega = 2\pi \cdot c / \lambda$

$\overrightarrow{E_{iz}}(x,t) = \overrightarrow{E_0} e^{i(kx-\omega t)}$





r

$\overrightarrow{E_e}(r,t) = -\frac{r_0}{r}\overrightarrow{E_0} e^{i(kr-\omega t)}$

(polarisation effects not indicated)

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

*r_0 is the **classical electron radius***

Incoming Flux Density :

$\phi_0\ (W/m^2) = \frac{1}{2}\ \varepsilon_0 \cdot c \cdot E_0^2$

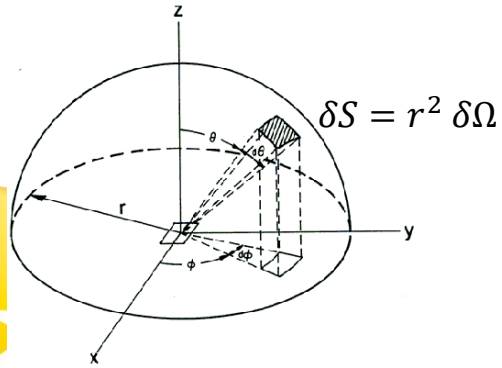
ϕ_e : **Flux Density scattered** by one (isolated) e^- at distance r

$\phi_e\ (W/m^2) = \frac{1}{2}\ \varepsilon_0 \cdot c \cdot E_e^2 = \frac{r_0^2}{r^2}\ \phi_0^2$

δF_e : flux scattered by one (isolated) e^- through a surface, $\delta S = r^2\ \delta\Omega$

$\delta F_e(W) = \phi_e \cdot \delta S = r_0^2 \cdot \phi_0 \cdot \delta\Omega$

The scattered flux only depends on the scattering solid angle and r_e, not on r



$\delta S = r^2\ \delta\Omega$

$\sigma = \frac{\delta F}{\delta\Omega} / \phi_0$ is called the ***differential scattering cross section***

$\sigma_e = r_0^2$ *is the electron differential scattering cross section*

Scattering amplitude by N electrons

Coherent scattering : summing up amplitudes

- Waves scattered by two electrons

Electron 1

$$\overrightarrow{E_{s1}}(r,t) = -\frac{r_0}{r}\overrightarrow{E_0}e^{i(kr-\omega t)}$$

Electron 2

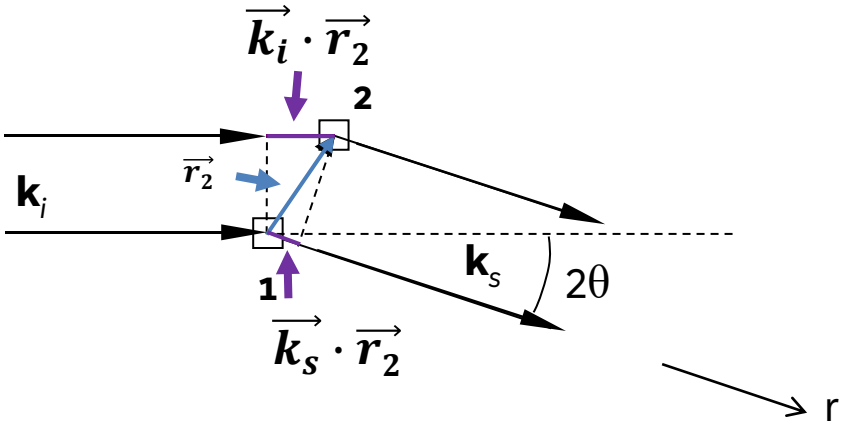
$$\overrightarrow{E_{s2}}(r,t) = -\frac{r_0}{r}\overrightarrow{E_0}e^{i(kr-\omega t+\overrightarrow{k_i}\cdot\overrightarrow{r_2}-\overrightarrow{k_s}\cdot\overrightarrow{r_2})}$$

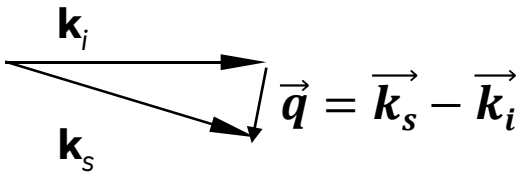
Path shift between waves 1 and 2 :

$$\overrightarrow{k_i}\cdot\overrightarrow{r_2} - \overrightarrow{k_s}\cdot\overrightarrow{r_2} = (\overrightarrow{k_i} - \overrightarrow{k_s})\cdot\overrightarrow{r_2} = -\vec{q}\cdot\overrightarrow{r_2}$$

$$\overrightarrow{E_{s2}}(r,t) = -\frac{r_0}{r}\overrightarrow{E_0}e^{i(kr-\omega t)}\, e^{-i\vec{q}\cdot\overrightarrow{r_2}}$$

$$\overrightarrow{E_{s1\&s2}}(r,t) = -\frac{1}{r}\overrightarrow{E_0}e^{i(kr-\omega t)}\, r_0\left(1+e^{-i\vec{q}\cdot\overrightarrow{r_2}}\right)$$





$\vec{q}$ = Momentum transfer

$$q = \frac{4\pi\sin(\theta)}{\lambda}$$

- Wave scattered by N electrons

$$\overrightarrow{E_{total}}(r,t) = -\frac{1}{r}\overrightarrow{E_0}e^{i(kr-\omega t)}\, r_0 \sum_{j=1}^{N} e^{-i\vec{q}\cdot\overrightarrow{r_j}}$$

$$\overrightarrow{E_{total}}(r,t) = -\frac{1}{r}\overrightarrow{E_0}e^{i(kr-\omega t)} A(\vec{q})$$

Scattering « amplitude » (length)

$$A(\vec{q}) = -r_0 \sum_{j=1}^{N} e^{-i\vec{q}\cdot\overrightarrow{r_j}}$$

Scattering amplitude from an atom

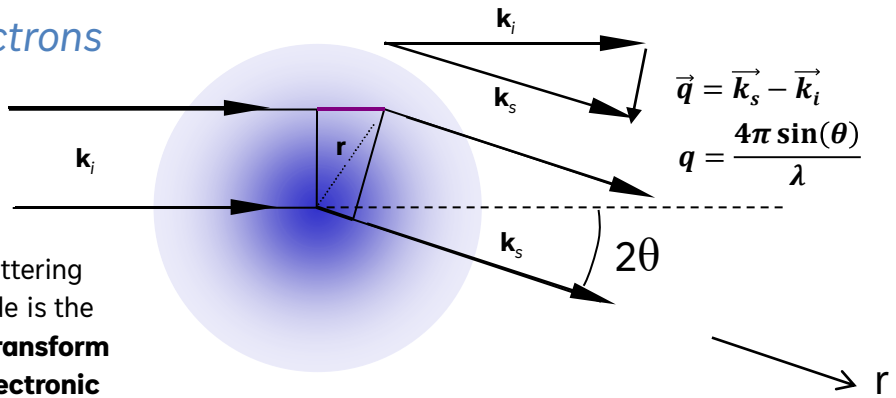
Scattering *amplitude from a particle with N electrons*

discrete representation

$$A(\vec{q}) = -r_0 \sum_{j=1}^{N} e^{-i\vec{q}\cdot\vec{r_j}}$$

continuous representation

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



The scattering amplitude is the **Fourier Transform of the electronic scattering length density**.

Scattering *amplitude from an atom*

In a atom, the electron cloud density $\rho_e(r)$ is purely radial

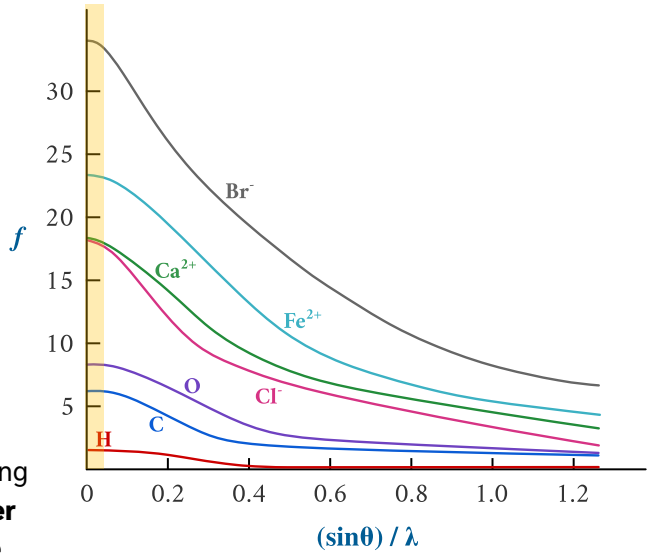
$$A_{at}(q) = -4\pi\, r_0 \int_0^{R_{at}} dr\, \rho_e(r) r^2 \frac{\sin(qr)}{qr}$$

Atomic scattering factor

$$f_{at}(q) = 4\pi \int_0^{R_{at}} dr\, \rho_e(r) r^2 \frac{\sin(qr)}{qr}$$

The atomic scattering factor is the **Fourier Transform of the electronic density** of a given type of atom.

$$A_{at}(q) = -r_0 f_{at}(q)$$



$A_{at}(q)$ and $f_{at}(q)$ are real (not complex)

$f_{at}(0) = Z$

Scattering from a molecule

Scattering *amplitude from a molecule with N atoms*

As a good approximation, the electrons pertaining to a given atom are considered to be positionned at its center. The **atomic scattering factor f_j** then replaces the sum of the electronic contributions from that atom in the calculation of the **molecule scattering amplitude**.

$$A_{molec}(\vec{q}) = -r_0 \sum_{j=1}^{N} f_j(q) e^{-i\vec{q}\cdot\vec{r_j}}$$

Scattering *intensity from a molecule with N atoms*

The scattering intensity is the square of the amplitude modulus.

$$I(\vec{q}) = AA^*(\vec{q})$$

$$I(\vec{q}) = {r_0}^2 \sum_{j=1}^{N}\sum_{k=1}^{N} f_j(q)\, fk(q)\, e^{-i\vec{q}\cdot(\vec{r_j}-\vec{r_k})}$$

The scattering intensity is the equivalent of a scattering differential cross-section.

If randomly oriented

Averaging over all orientations:

$$\langle e^{-i\vec{q}\cdot r}\rangle = \frac{\sin(qr)}{qr}$$

$$I(q) = {r_0}^2 \sum_{j=1}^{N}\sum_{k=1}^{N} f_j(q)\, fk(q) \frac{\sin(qrij)}{qr_{ij}}$$

Debye formula

Scattering from a molecule

Scattering *amplitude*

$$A(\vec{q}) = -r_0 \int_{V_{molec}} \rho_e(\vec{r}) e^{-i\vec{q}\cdot\vec{r}} d^3\mathbf{r} = -r_0\, FT[\rho_e(\vec{r})]$$

Scattering *intensity*

$$I(\vec{q}) = A.A^*(\vec{q}) = {r_0}^2\, FT[\rho_e(\vec{r})]\, FT[\rho_e(-\vec{r})] = {r_0}^2 FT[\rho_e(\vec{r}) * \rho_e(-\vec{r})]$$

$$\gamma_e(\vec{r}) = \int_{V'} \rho_e(\vec{r} + \vec{r}\,')\rho_e(\vec{r'})\, d^3\mathbf{r}'$$

$\gamma_e(\vec{r})$: electronic density autocorrelation function.

$$I(\vec{q}) = {r_0}^2 \int_V \gamma_e(\vec{r})\, e^{-i\vec{q}\cdot\vec{r}} d^3\mathbf{r}$$

The scattering intensity is the **Fourier Transform of the electronic density autocorrelation function** (times r_0^2).

If randomly oriented

Averaging over all orientations:

$$\gamma_e(r) = \langle \gamma_e(\vec{r}) \rangle$$

$$\mathrm{I(q)} = 4\pi\, {r_0}^2 \int_V \gamma_e(r) r^2 \frac{\sin(qr)}{qr} dr$$

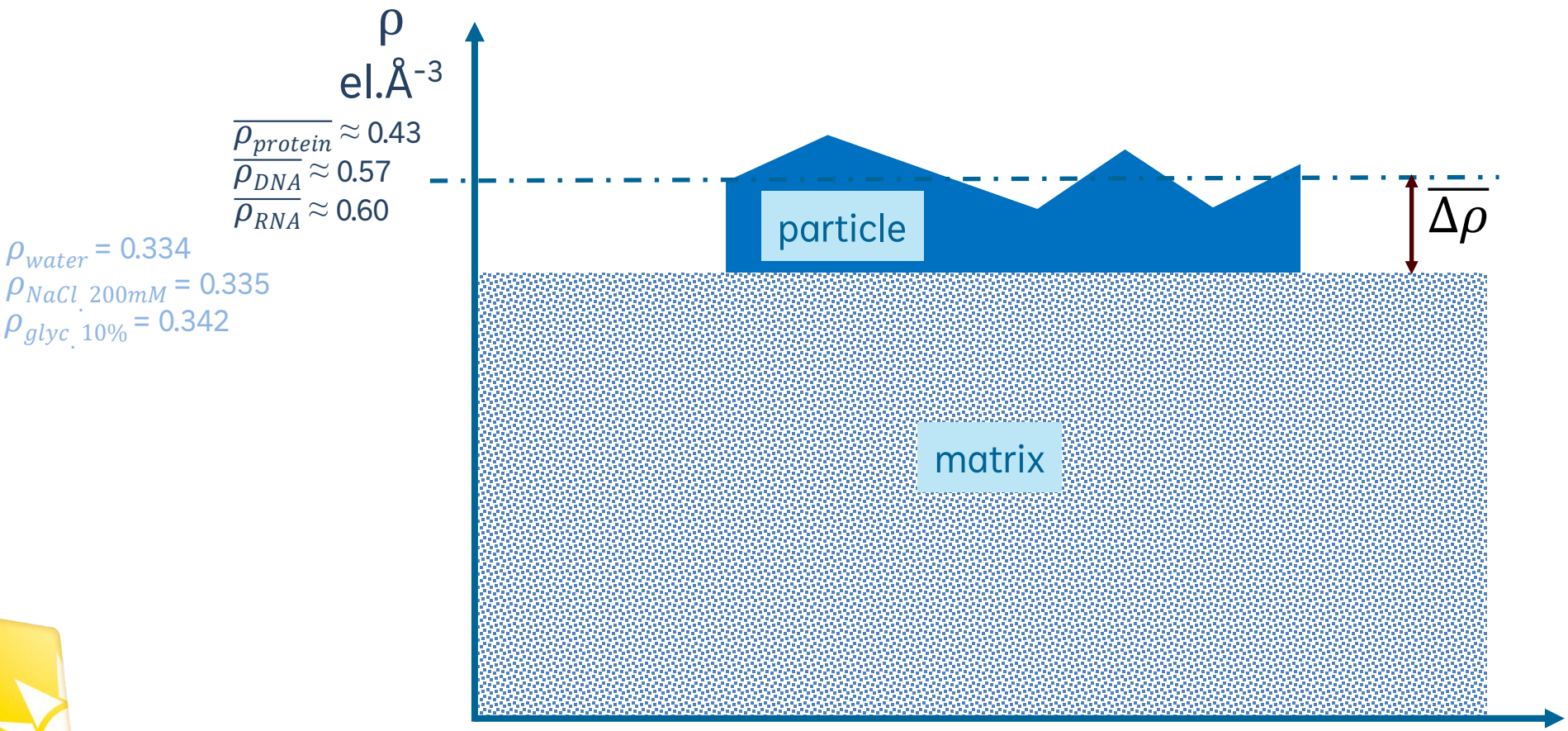
$$\gamma_e(r) = \langle {\rho_e}^2 \rangle V\, \gamma_0(r)$$

$\gamma_0(r)$ is the characteristic funtion

« probability of finding a point within the particle at a distance r from a given point »

Particles in a homogeneous matrix (or buffer)

- A particle is described by the associated electron density distribution $\rho_p(\vec{r})$.
- What contributes to scattering **at small angles** is the ***contrast*** of electron density between the particle and the matrix $\Delta\rho(\vec{r}) = \rho_p(\vec{r}) - \rho_0$, that is **very small** for biological samples.

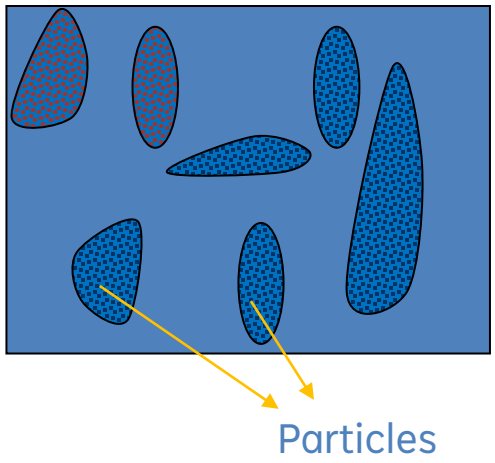


Particles in a homogeneous matrix (or buffer)

- Scattering amplitude at small angles

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

$$F(\vec{q}) = -r_0 \int_{V_{particles}} \Delta\rho_e(\vec{r})\, e^{-i\vec{q}\cdot\vec{r}} d^3\mathbf{r}$$



- $\Delta\rho_e(\vec{r})$ is the contrast of electronic density and describes the scattering objects
- $F(\vec{q})$ is the Scattering Amplitude at small angles of the ensemble of particles

- Scattering intensity per unit volume

$$I(\vec{q}) = \frac{1}{V_{sample}} F(\vec{q}) F^*(\vec{q})$$

- $I(\vec{q})$ 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 directional correlations). The sample as a whole is therefore **isotropic**. As a result, the scattering intensity only depends on the **modulus** of $\vec{q}$, q = 4π sin(θ) /λ.

$$I(q) = \frac{1}{V_{sample}} \overline{\langle F(\vec{q}) F^*(\vec{q}) \rangle}$$

Modulus — Vector

If the solution is "**ideal**" (No correlations between particles positions = No short-range or long-range interactions), then **the individual intensities sum up**.

$$I(q) = \sum_{i=1,N} I_i(q) = \frac{1}{V_{sample}} \sum_{j=1,N} \overline{\langle F_j(\vec{q}) F_j{}^*(\vec{q}) \rangle_\Omega}$$

The averaged scattering intensity of a particle in an ideal solution is called its **form factor**, P(q).

$$\mathrm{P_j(q)} = \overline{\langle F_j(\vec{q}) F_j{}^*(\vec{q}) \rangle_\Omega}$$

$$I(q) = \frac{1}{V_{sample}} \sum_{i=1,N} P_i(q)$$

$$\mathrm{P_j}(0) = r_0{}^2 V^2_{particle\ j} \langle \Delta\rho_j \rangle^2$$

→ Average Electronic Density contrast

Particles in solution

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

$$I(q) = \frac{1}{V_{sample}} \overline{\langle F(\vec{q}) F^*(\vec{q}) \rangle}$$

Modulus — Vector

If the solution is "**ideal**" (No correlations between particles positions = No short-range or long-range interactions), then **the individual intensities sum up**.

$$I(q) = \sum_{i=1,N} I_i(q) = \frac{1}{V_{sample}} \sum_{j=1,N} \overline{\langle F_j(\vec{q}) {F_j}^*(\vec{q}) \rangle_\Omega}$$

If the solution is "**ideal" and "monodisperse**" (all particles are identical), then **the individual form factors are all identical**.

$$I_i(q) = I_{part}(q) \quad and \quad P_i(q) = P_{part}(q), \qquad whatever\ \mathrm{i}$$

$$I(q) = \frac{1}{V_{sample}} \sum_{i=1,N} P_i(q) = \frac{1}{V_{sample}} N\, \mathrm{P_{part}(q)} = \varphi\, \mathrm{P_{part}(q)}$$

Monodispersity and ideality

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

- **Ideality**
  - Yes ← No correlations between particles positions

    (No long-range interactions)
  - No ← Correlations between particles positions

    (Existence of short-range or long-range interactions)

- **Polydisperse and Ideal**

$$I(q) = \sum_{i=1,N} I_i(q) = \frac{1}{V} \sum_{i=1,N} P_i(q)$$

- **Polydisperse and Not ideal**

$$I(q) = \left(\frac{1}{V} \sum_{i=1,N} P_i(q) \right) \cdot S(q)$$

S(q) is called the **Structure Factor** (of the sample)
S(q) = 1 for an ideal solution
S(q) usually may differ from 1 at very small q values

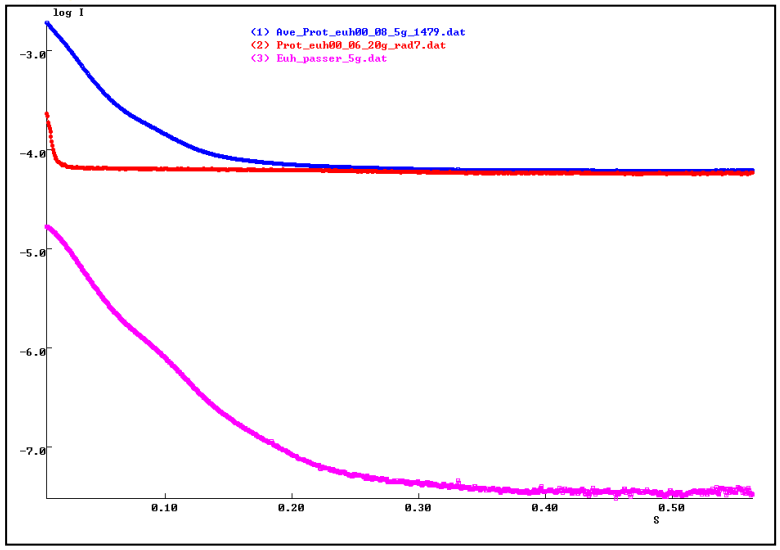
- **Monodisperse but not ideal**

$$I(q) = \frac{\mathrm{N}}{V} P_{part}(q)\ \mathrm{S(q)}$$

- **Monodisperse ideal**

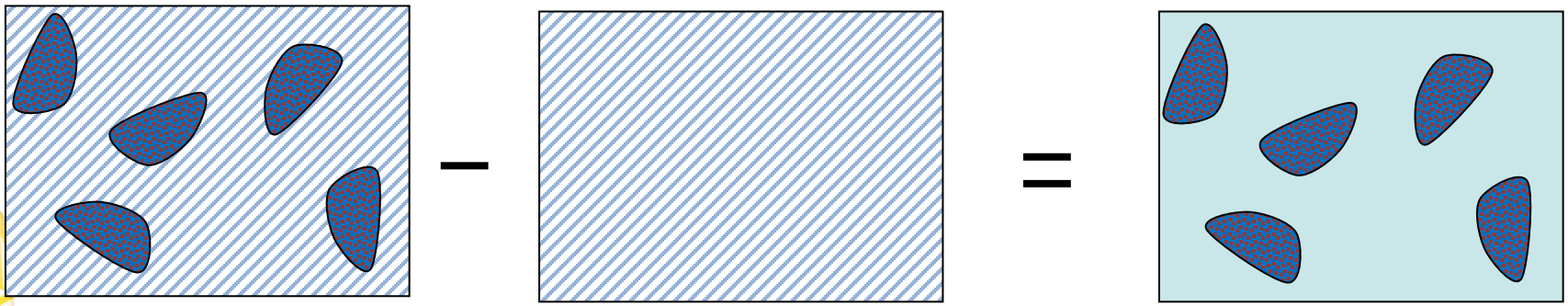
$$I(q) = \mathrm{N}\, I_{part}(q) = \frac{\mathrm{N}}{V} P_{part}(q)$$

Solvent scattering and contrast



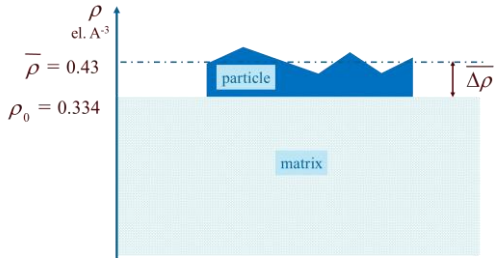
To obtain scattering solely from the contrasting particles, intrinsic solvent scattering must be measured **very accurately** and subtracted, which also permits to subtract contribution from parasitic background (slits, sample holder etc).

$$I_{solution}(q) - I_{buffer}(q) = \boldsymbol{I_{particles}(q)}$$



Do not confuse !

contrast effect

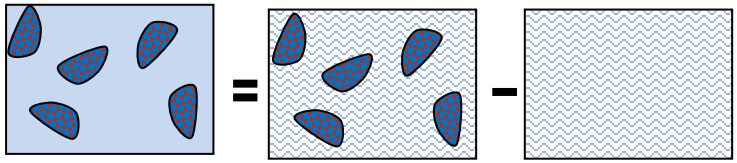


$$\Delta\rho_e(\vec{r}) = \rho_e(\vec{r}) - \rho_0$$

$$F(\vec{q}) = -r_0 \int_{V_{particles}} \Delta\rho(\vec{r})\, e^{-i\vec{q}\cdot\vec{r}}\, d^3\mathbf{r}$$

$$I_{particles}(q) = \frac{1}{V_{sample}} \overline{\langle F(\vec{q}) F^*(\vec{q}) \rangle_\Omega}$$

buffer subtraction



$$I_{particles,\,exp}(q) = I_{solution,\,exp}(q) - I_{buffer,\,exp}(q)$$

BIOPHYSICAL INFORMATION

Biophysical information

- Guinier Analysis
- Kratky plot : why is it so interesting ?
- « Real-space SAXS » : Pair distribution function P(r)

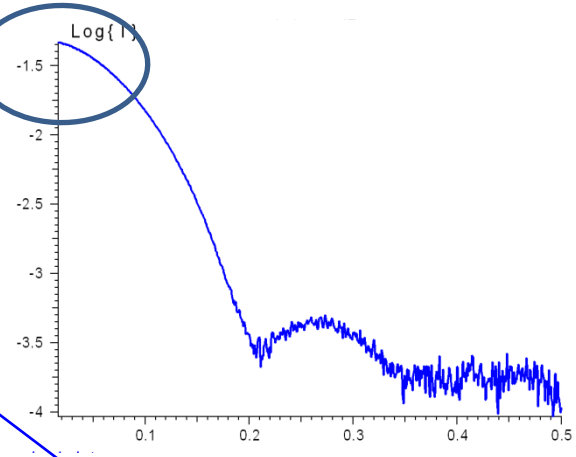
Biophysical information

- Guinier Analysis
- Kratky plot : why is it so interesting ?
- « Real-space SAXS » : Pair distribution function P(r)

Asymptotic behaviour at small angles : 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 / Rg for a globular shape.



Prof. André Guinier
1911-2000
Orsay, France

$$I(q) = I(0)\exp\left(\frac{-q^2Rg^2}{3}\right)$$

Extrapolated intensity at origin

Radius of gyration

Guinier law, in Log scale :

$$Ln[I(q)] = Ln[I(0)] - \frac{q^2Rg^2}{3}$$

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 Rg and I(0).

Mass retrieval from Guinier analysis

$$I(q) = I(0)\exp\left(\frac{-q^2 Rg^2}{3}\right)$$

Absolute Unit : cm^{-1}

Classical electron radius

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

$$Rg^2 = \frac{\int_V r^2\left(\rho_{prot}(\vec{r}) - \rho_{buf}\right)d^3\vec{r}}{\int_V \left(\rho_{prot}(\vec{r}) - \rho_{buf}\right)d^3\vec{r}}$$

Mass concentration

Electronic density contrast

Protein specific volume

I(0) gives an independent estimation of the molar mass of the protein **(only if the mass concentration and specific volume are precisely known …)**

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

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

For unfolded proteins : $R_g\,(Å) \approx 8.05 * M^{0.522}$

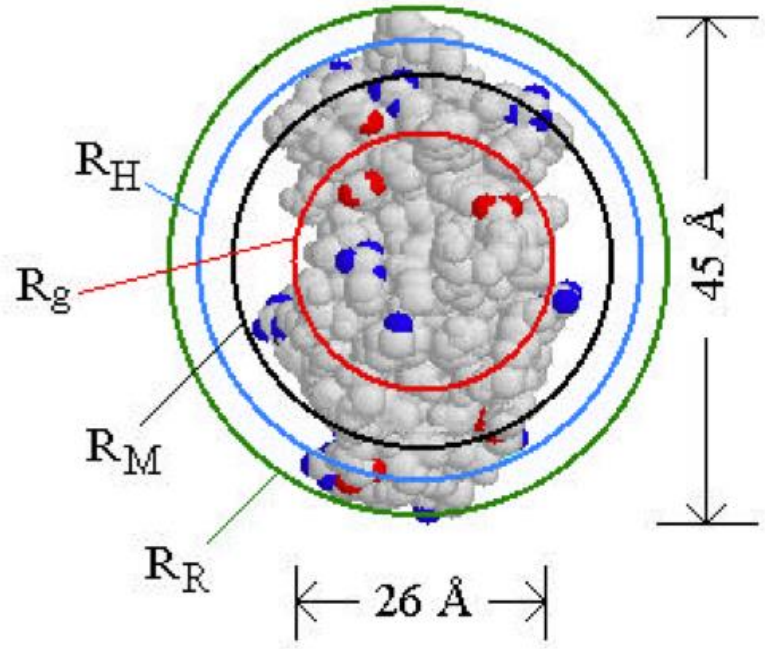
Bernado et al. (2009), Biophys. J., 97 (10), 2839-2845.

Typically :

M (kDa) = (1200 ~ 1600) * I0 (cm^{-1}) / C (mg/ml)

Radius of gyration

Lysozyme



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

$$Rg^2_{SAXS} = \frac{\int_V r^2 \Delta\rho_e(\vec{r})\, d\vec{r}}{\int_V \Delta\rho_e(\vec{r})\, d\vec{r}}$$

Useful definitions of R_g

$$R_g^2 = \frac{1}{N}\sum \left\| \vec{r_i} - \overrightarrow{r_{COM}} \right\|^2 \quad \text{by atoms}$$

$$R_g^2 = \int_V r(r) r^2\, dr \Big/ \int_V r(r)\, dr \quad \text{by electron density}$$

$$R_g^2 = \frac{1}{2N(N-1)} \sum_i \sum_j \left\| \vec{r_i} - \vec{r_j} \right\|^2 \quad \text{by atom pairs}$$

$$R_g^2 = \tfrac{1}{2} \int r^2 p(r)\, dr \Big/ \int p(r)\, dr \quad \text{by pair distribution}$$

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

* center of mass of the *electron* density

Sphere $R_g = \sqrt{\frac{3}{5}} R$

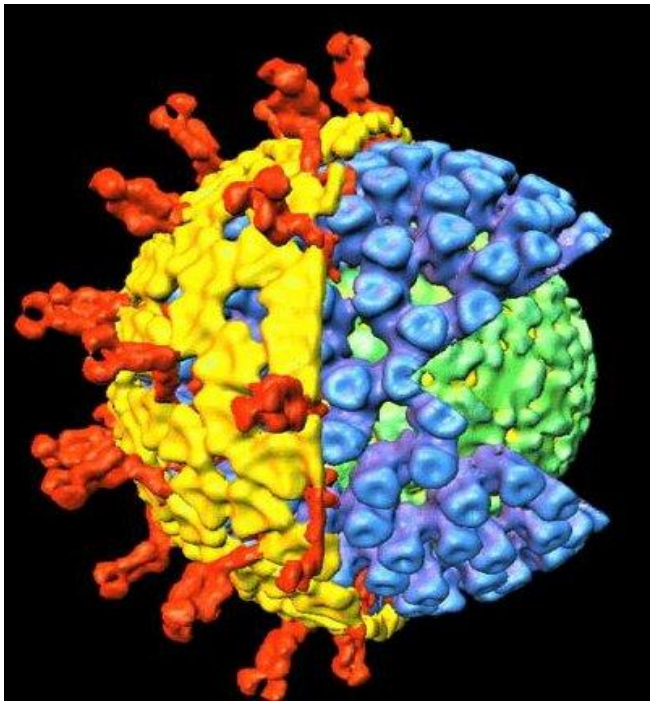
Thin rod $R_g = \sqrt{\frac{1}{12}} L$

Thin disk $R_g = \sqrt{\frac{1}{2}} R_{disk}$

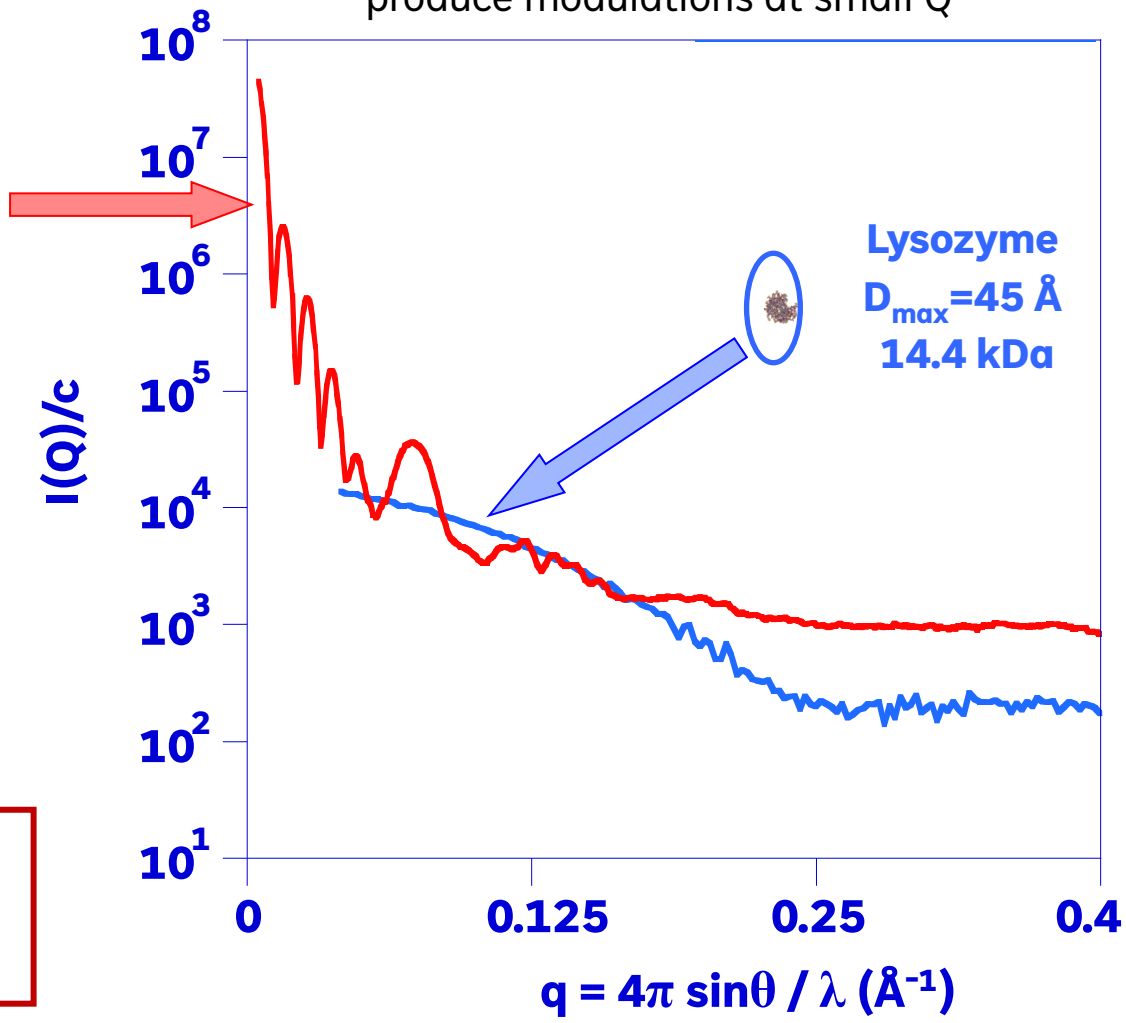
Courtesy: Richard Gillilan, Cornell U., USA

Radius of gyration

Rotavirus VLP : diameter = 750 Å, 44 MDa



Basic law of reciprocity in scattering:
Correlations in large objects
produce modulations at small Q



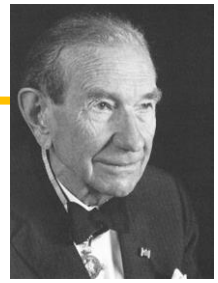
Rule of thumb for the needed q_{min} for a correct Guinier fit:

$q_{min} < 1 / D_{max}$

Biophysical information

- Guinier Analysis
- Kratky plot : why is it so interesting ?
- « Real-space SAXS » : Distance correlation function P(r)

Kratky Plot

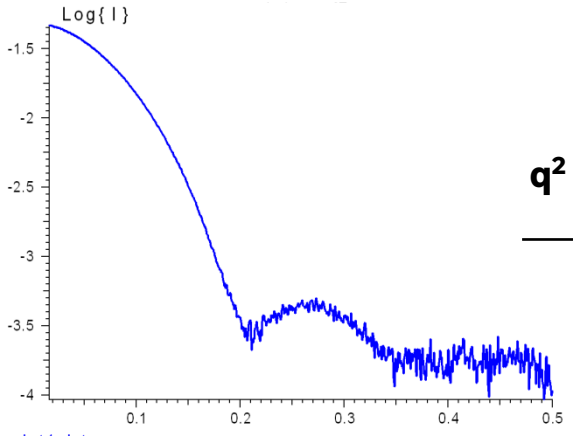


Prof. Otto Kratky
1902-1995
Graz, Austria

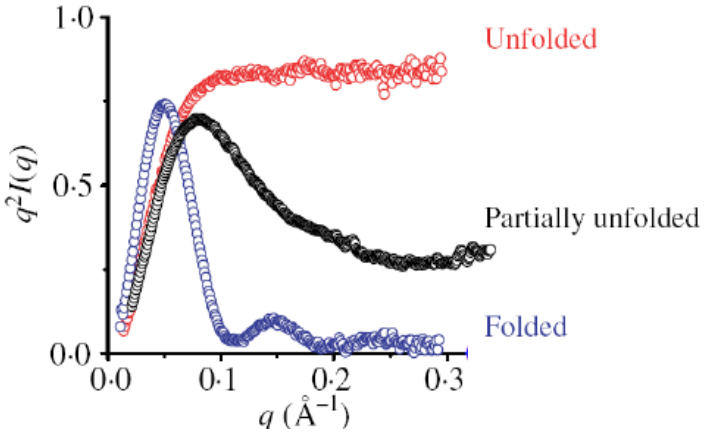
SAXS provides a sensitive means to ***evaluate the degree of compactness*** of a protein:

- 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:



q^2 I(q) *versus* q



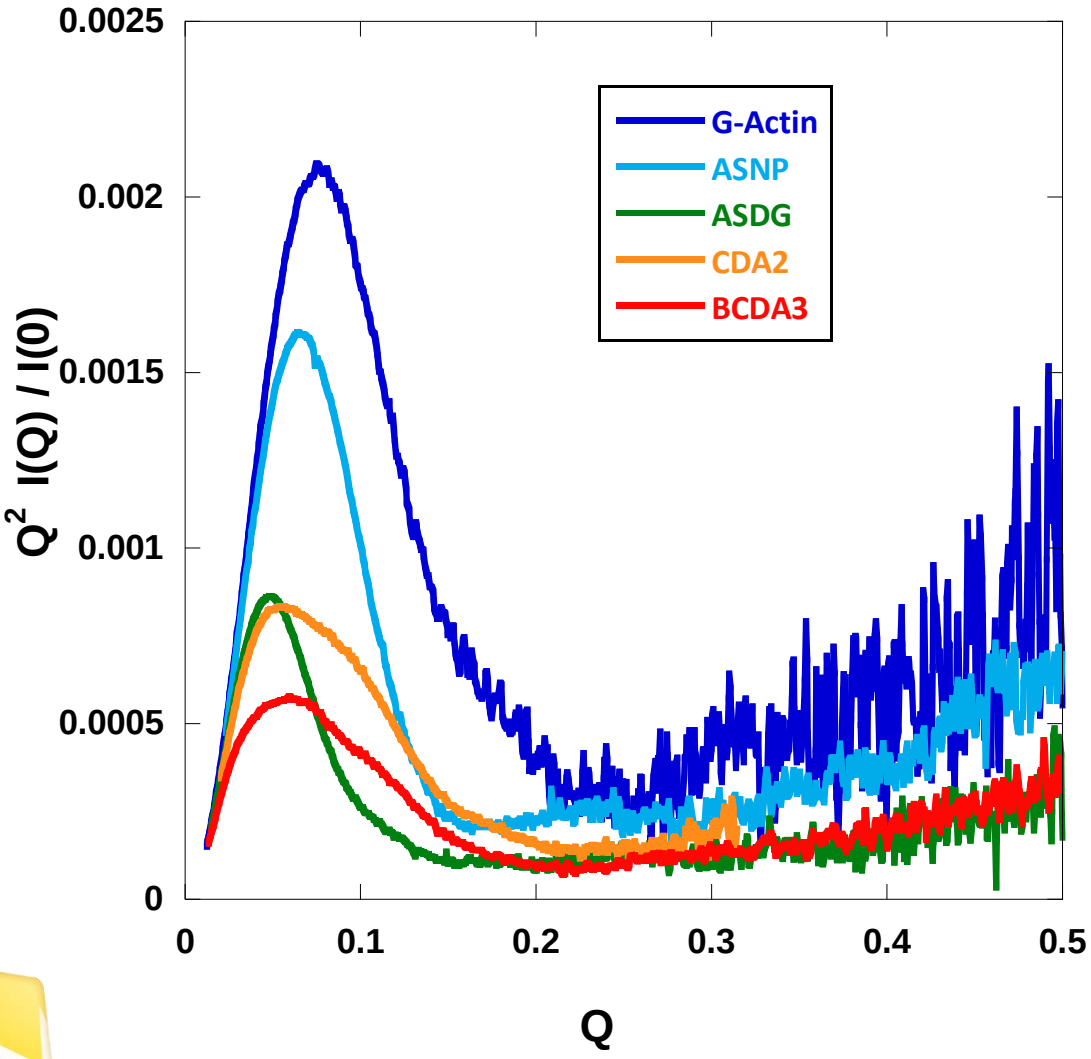
Putnam, D., et al. (2007) Quart. Rev. Biophys. 40, 191-285.

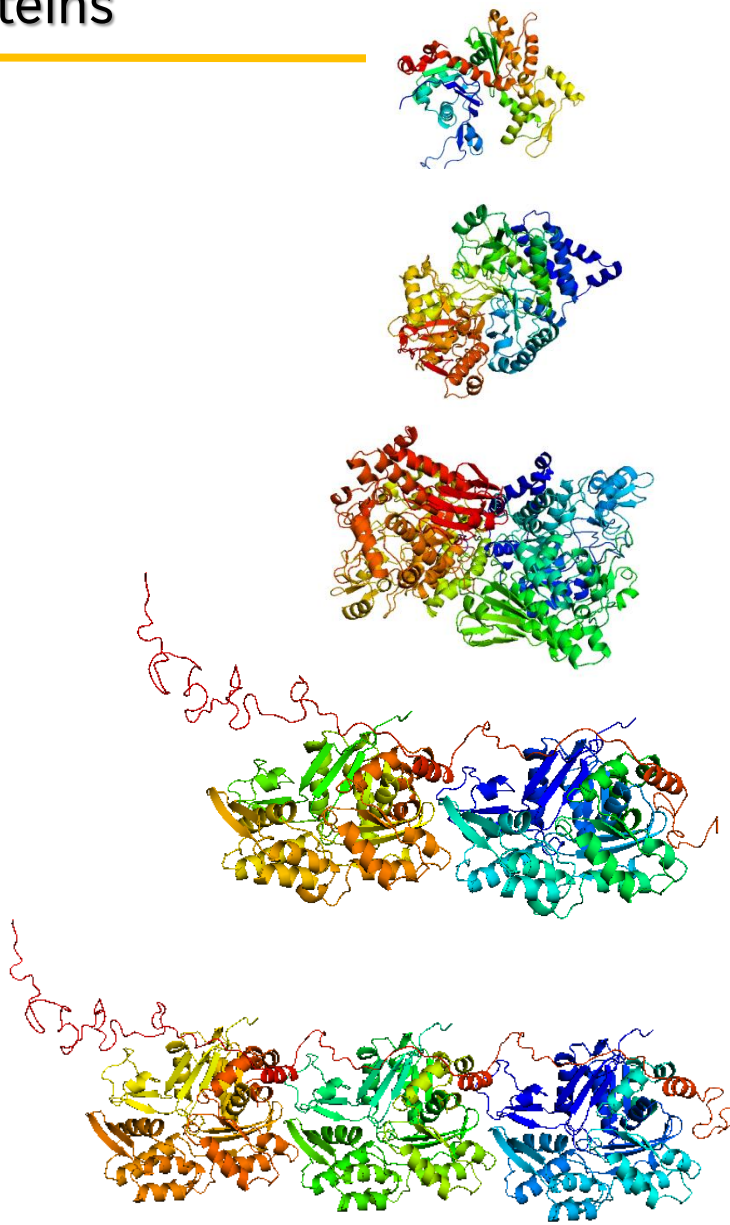
Folded particle : ***bell-shaped curve*** (asymptotic behaviour $I(q) \sim q^{-4}$)

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 $I(q) \sim q^{-1.x}$)

Kratky Plots of folded proteins





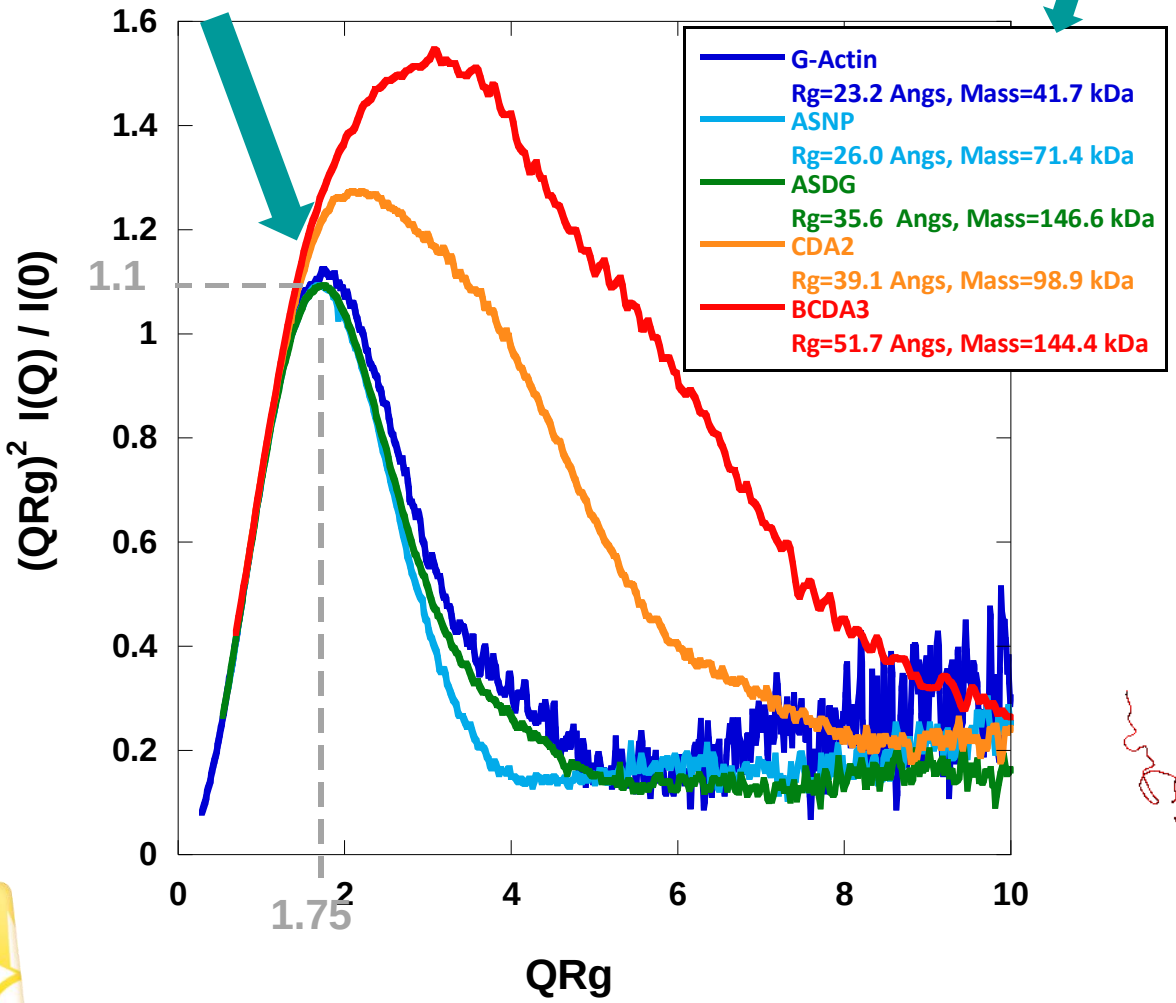
Folded proteins display a bell shape. Can we go further?

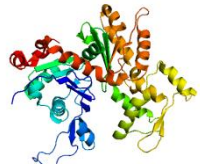
Dimensionless Kratky Plots of folded proteins

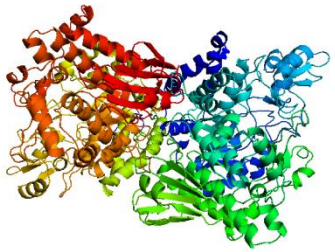
Introduced for biology in Durand et al. (2010), J. Struct. Biol. **169**, 45-53.

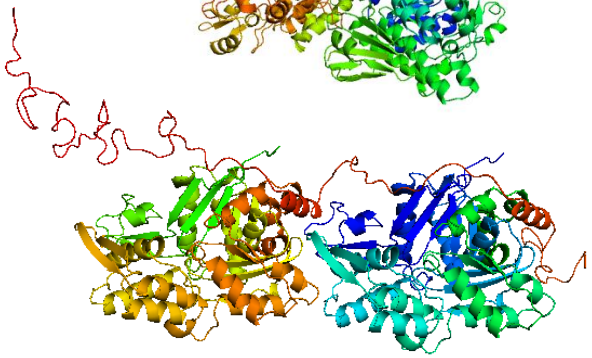
For globular structures, the plots fold into the same maximum

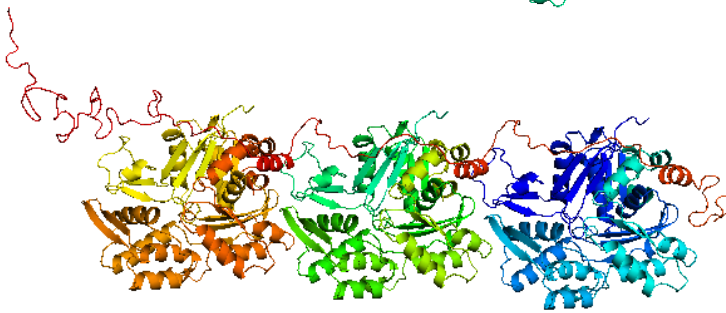
The relation $M_{Rg}(kDa) \approx (Rg / 6.5)^3$ only works for the globular structures, not the elongated







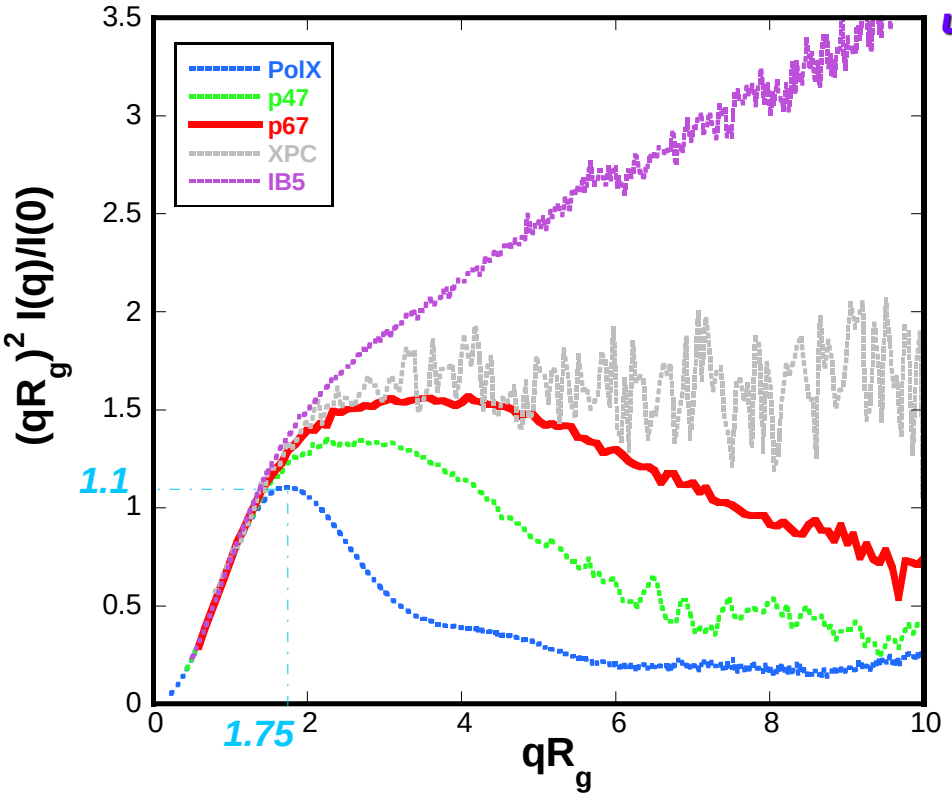




The position of the maximum on the dimensionless bell shape tells to what extent the protein is globular.

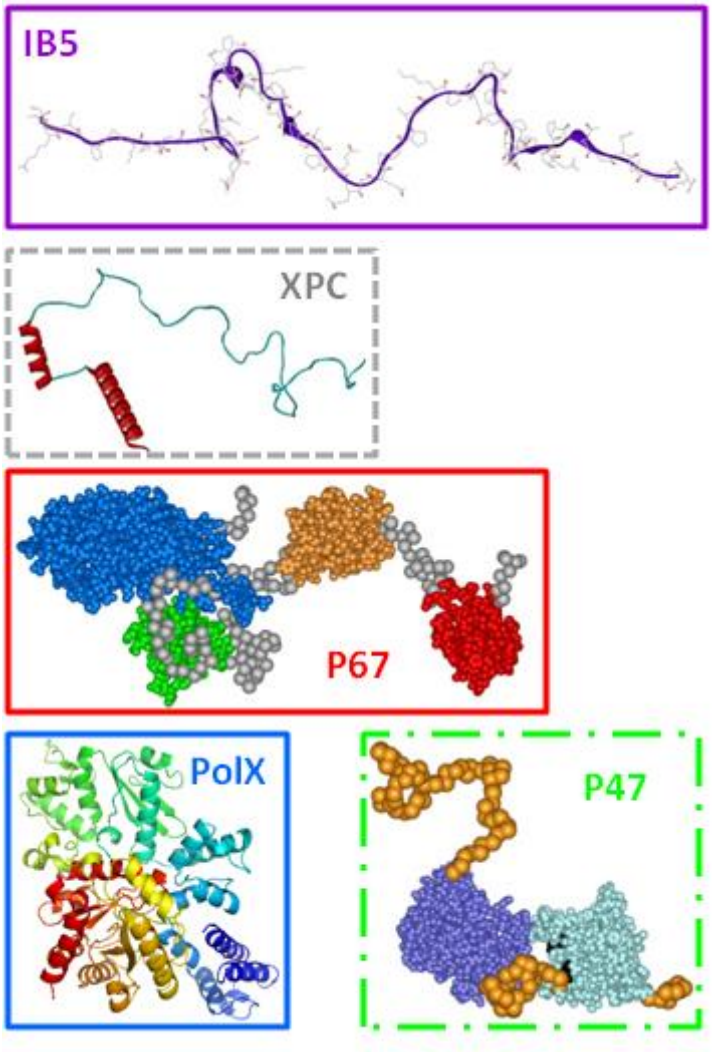
Dimensionless Kratky Plots of (partially) unfolded proteins

Receveur-Bréchot V. and Durand D (2012), Curr. Protein Pept. Sci., 13:55-75.



unfolded

globular



The bell shape vanishes as folded domains disappear and flexibility increases.

The curve increases at large q as the structure extends.

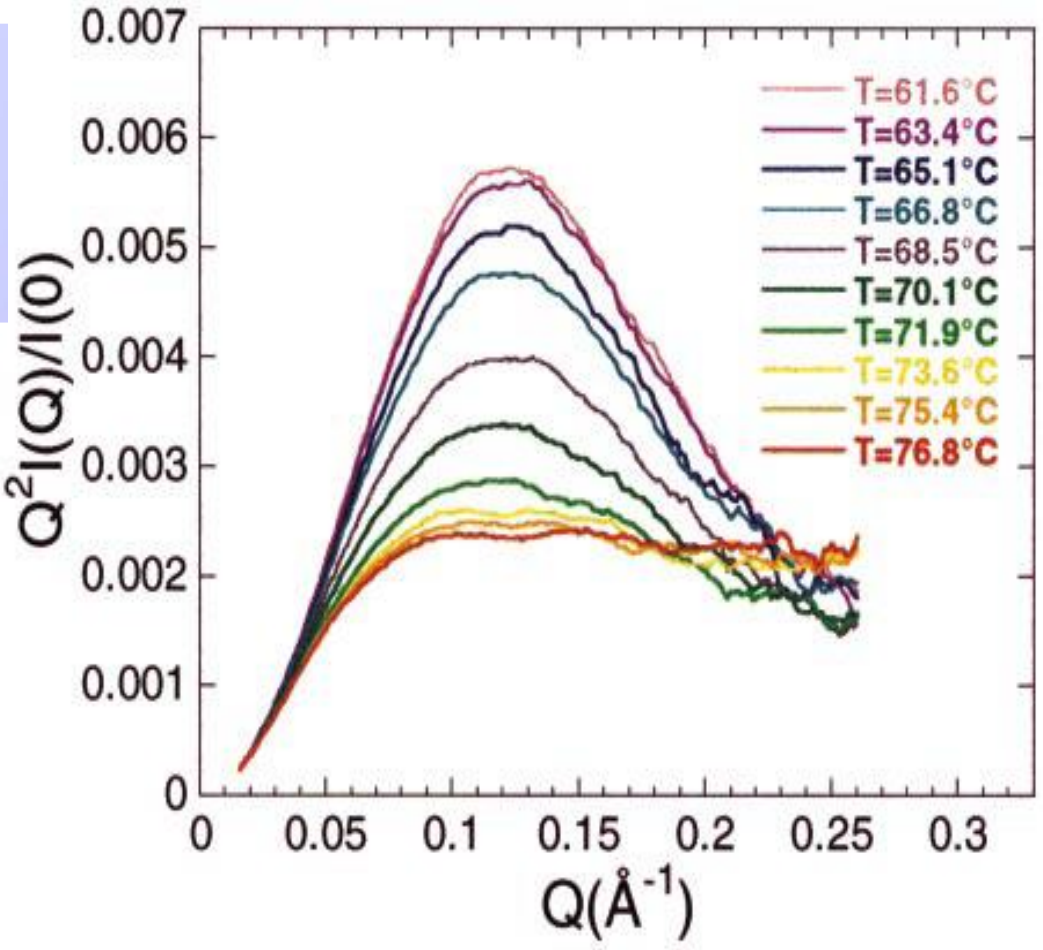
Kratky Plot : NCS heat unfolding

Pérez et al., *J. Mol. Biol.*(2001), 308, 721-743

T=22°C T=76°C

In practice, thin Gaussian chains do not exist.

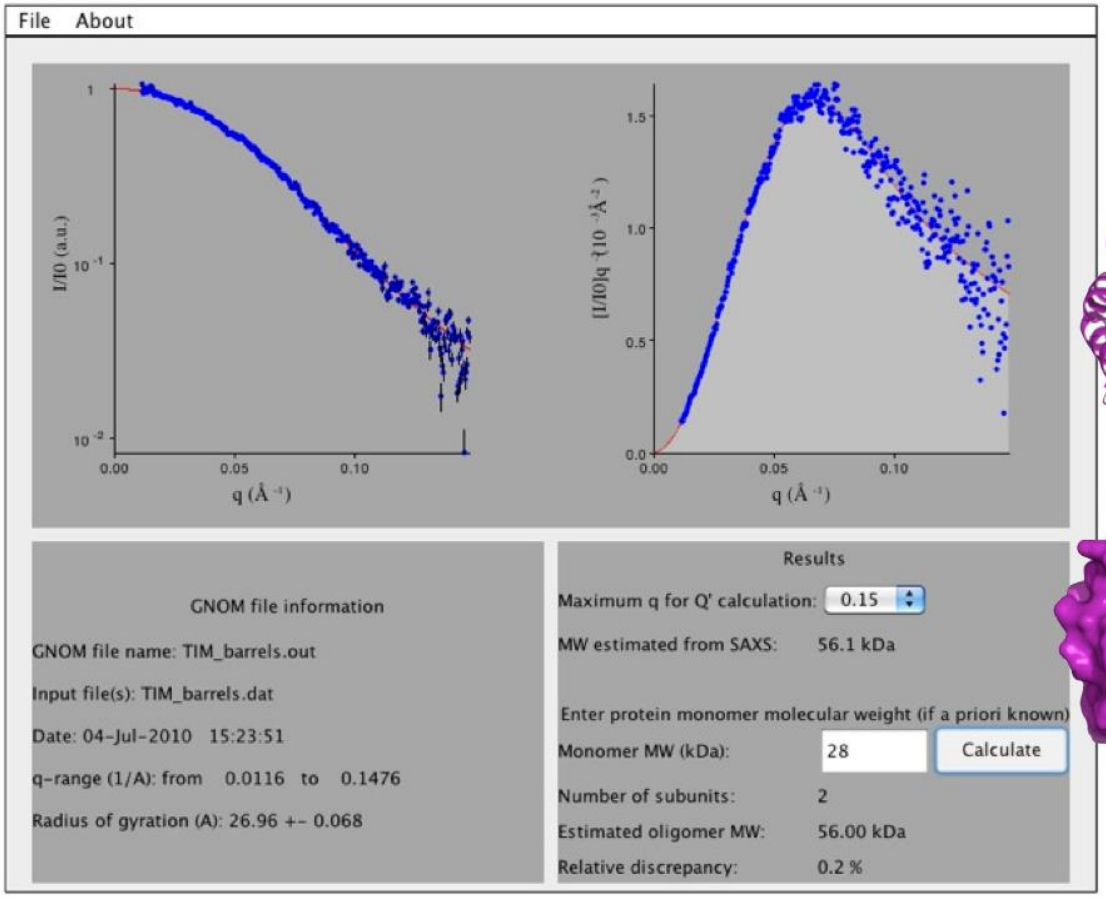
In spite of the plateau at T=76°C, NCS is not a Gaussian chain when unfolded, but a thick chain with persistence length

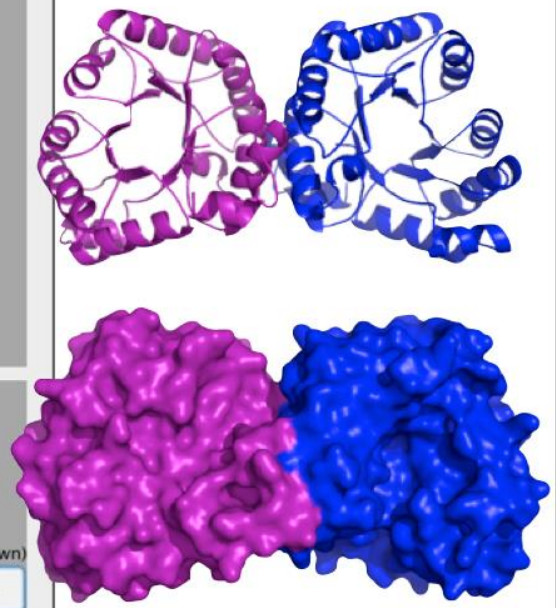


Molecular Weight estimation based on Porod invariant

http://saxs.ifsc.usp.br/

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





Other methods for MW estimation based on similar though different grounds were developed

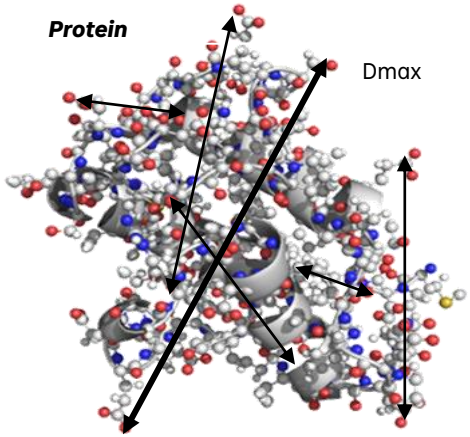
Rambo R. And Tainer J. (2013), Nature, **496**, 477-481.

Biophysical information

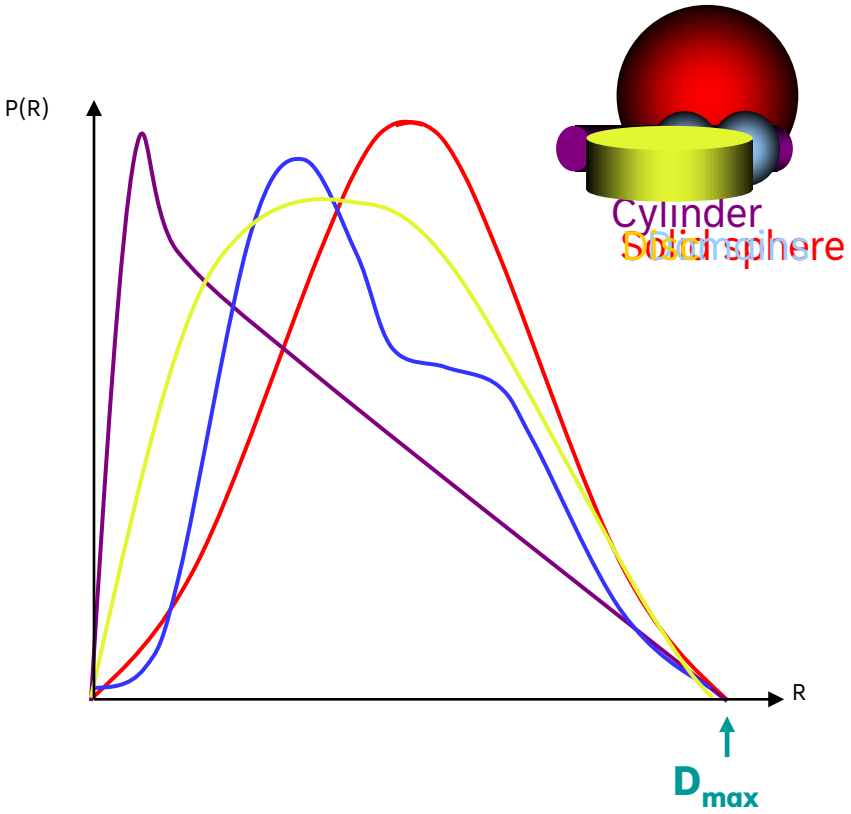
- Guinier Analysis
- Kratky plot : why is it so interesting ?
- « Real-space SAXS » : Pair distribution function P(r)

Pair Distribution Function p(r)

The pair distribution function p(r) is proportional to the average number of neigbouring atoms at a given distance, r, from any given atom within the macromolecule.



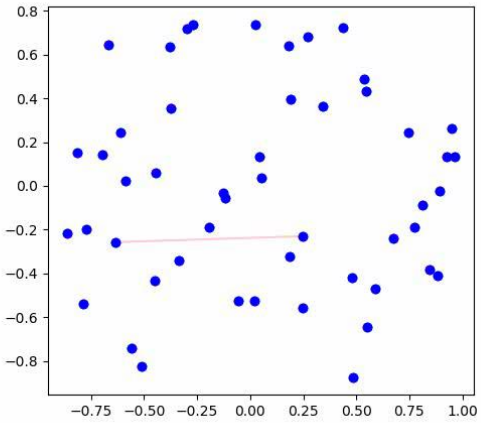
p(r) vanishes at r = D_{max}

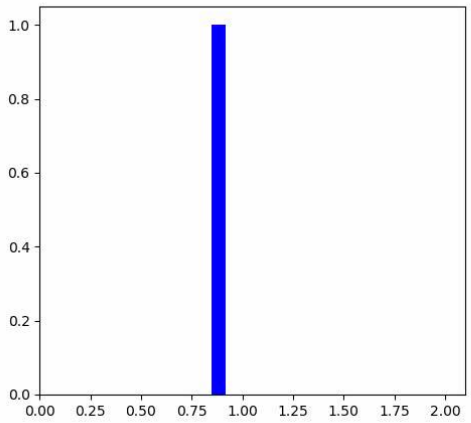


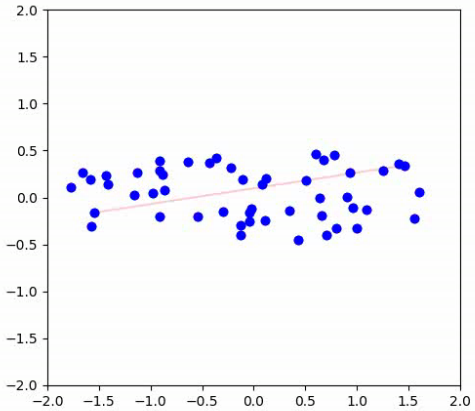
The distance distribution function characterises the shape of the particle **in real space**

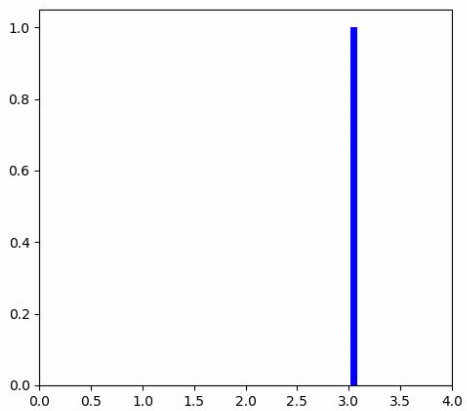
Pair Distribution Function p(r)

Marc-André Delsuc:
thanks for the python script









Relation between p(r) and I(Q)

The intensity is related to the Fourier Transform of the self-correlation function $\gamma_{obj}(r)$

$$I(q) = 4\pi\, {r_e}^2 \varphi \int_{V_{obj}} \gamma_{obj}(r) r^2 \frac{\sin(qr)}{qr} dr$$

Fourier Transform for isotropic samples

And the pair distribution function is directly related to the self-correlation function

$$p(r) = \gamma_{obj}(r) r^2$$

Then :

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

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

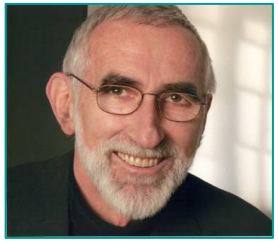
Both curves contain the same information.

Apparently, p(r) could be directly derived from I(q).

However, direct calculation of p(r) from I(q) is made difficult and risky because of $[q_{min}, q_{max}]$ truncation and data noise effects.

Back-calculation of the Distance Distribution Function

Glatter, O. *J. Appl. Cryst.* (1977) **10**, 415-421.



Prof. Otto Glatter
Guinier Prize 2012
Graz, Austria

Main hypothesis : the particle has a « finite » size, characterised by D_{max}.

- D_{max} is proposed by the « user »
- A guess for p(r) is decomposed over [0, D_{Max}] by a linear combination of orthogonal functions

$$p_{calc}(r) = \sum_{1}^{M} c_n \, \phi_n(r)$$

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

$$I(q) = 4\pi \, r_e{}^2 \phi \int_0^{D_{\max}} p_{calc}(r) \frac{\sin(q \cdot r)}{q \cdot r} dr$$

- $\{c_n\}$ are optimized recursively

Dr. Dmitri Svergun
Guinier Prize 2018
Hamburg, Germany

Svergun (1988) : program "GNOM"

M ~ 30 - 100 ⇒ ill-posed LSQ ⇒ regularisation method

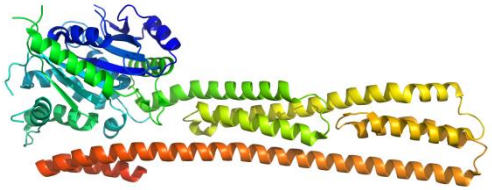
+ "Perceptual criteria" : smoothness, stability, absence of systematic deviations

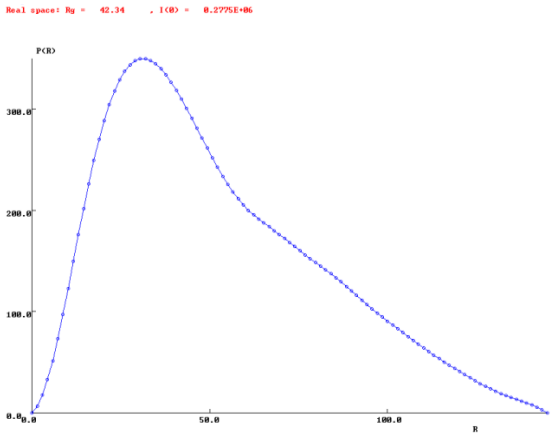
- Each criterium has a predefined weight
- The solution is given a score calculated by comparison with « ideal values »

Pair Distribution Function

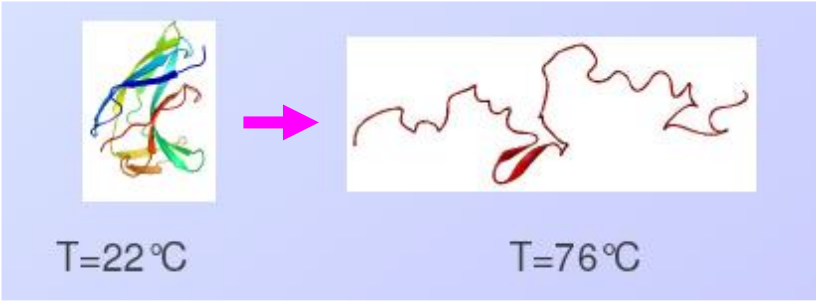
Experimental examples

GBP1

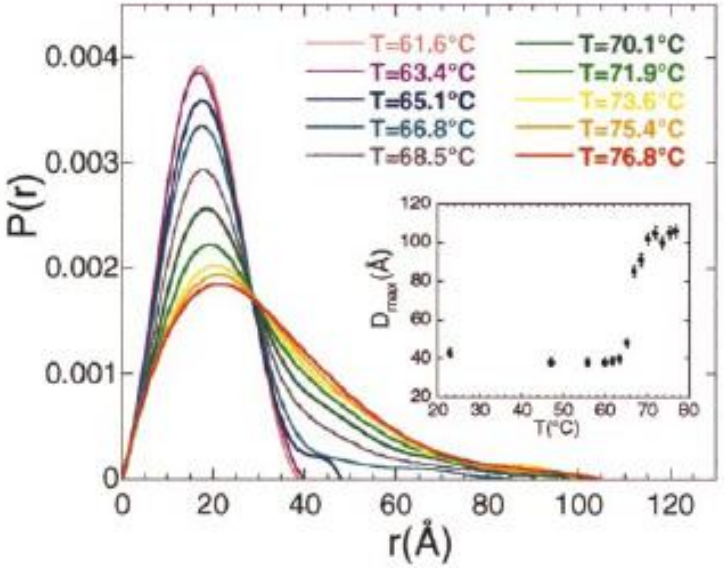




Heat denaturation of Neocarzinostatin

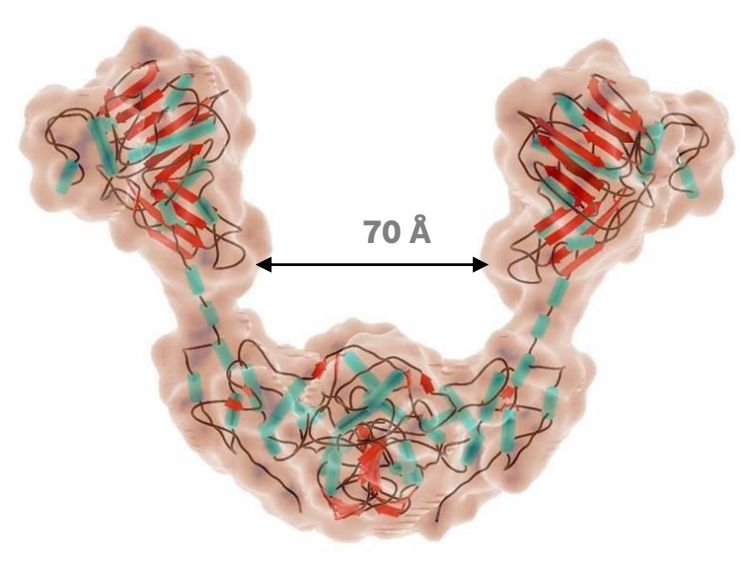


Pérez et al., J. Mol. Biol. (2001) 308, 721-743

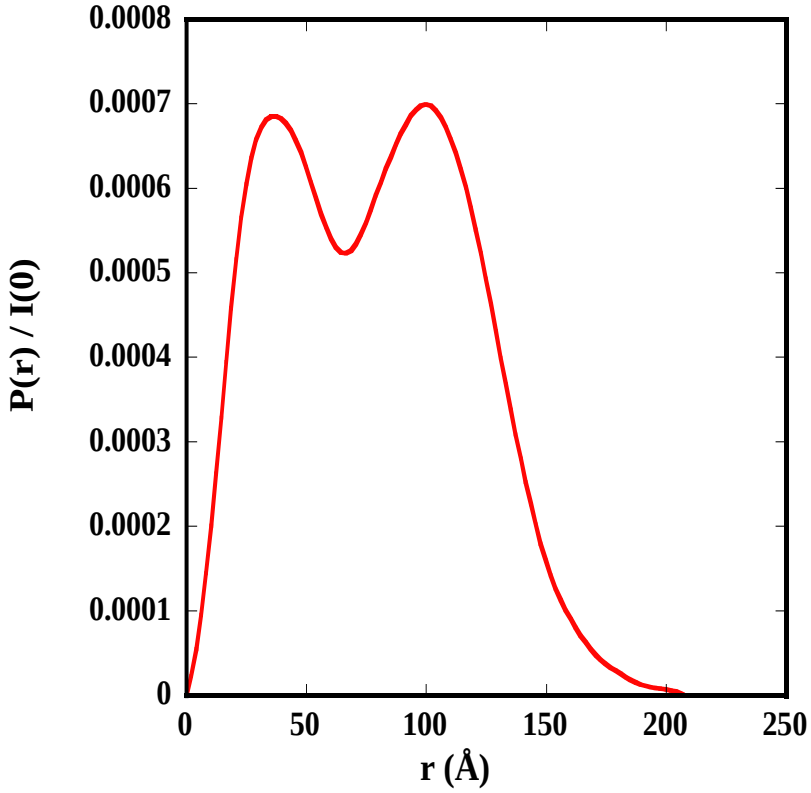


Experimental examples

Topoisomerase VI



Bimodal distribution

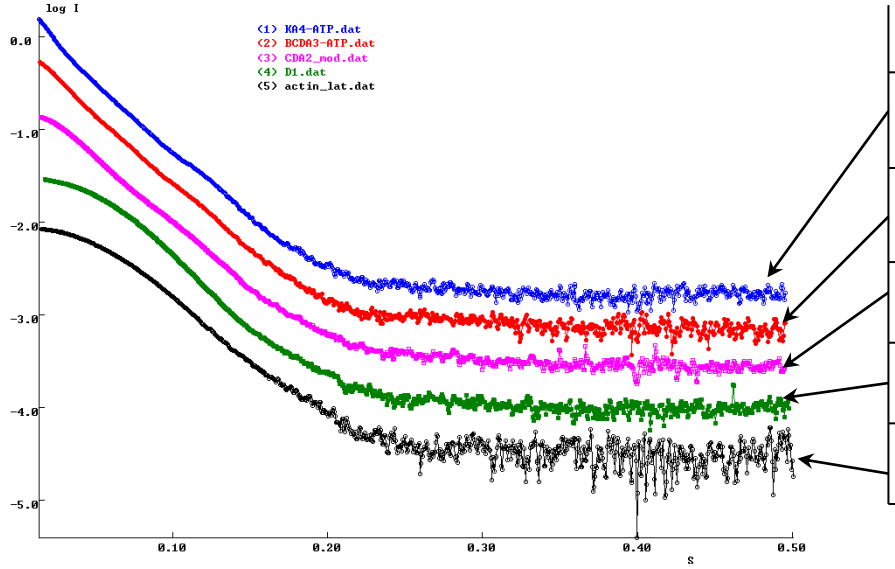


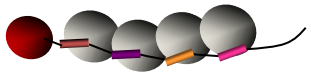
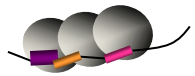
M. Graille et al., Structure (2008), *16*, 360-370.

Distance Distribution Function

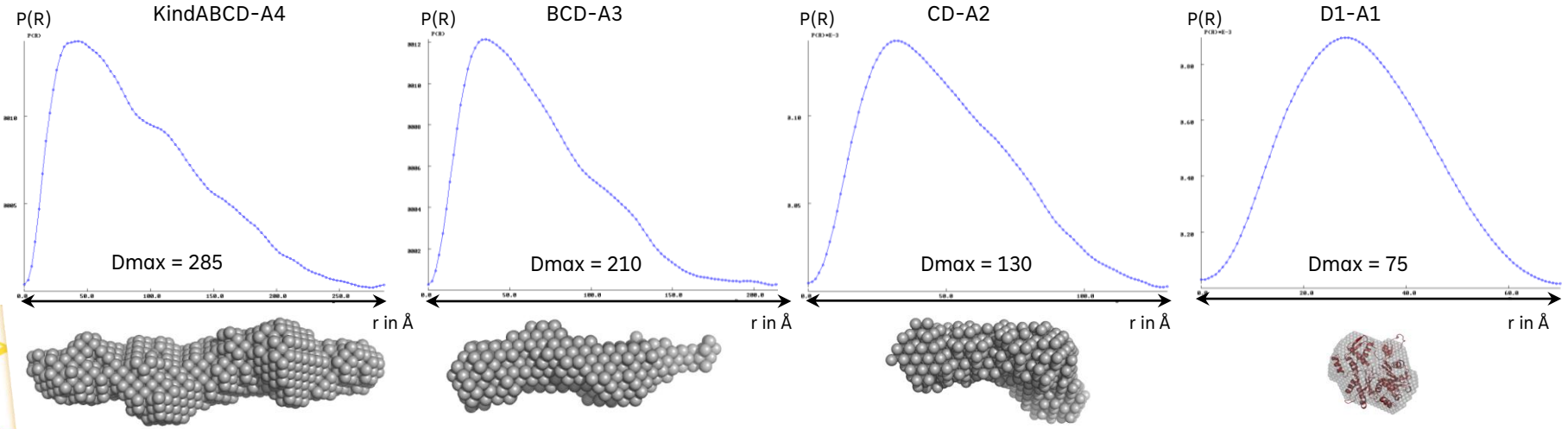
Scattering curves obtained on different complexes Spire-Actin and Actin alone

Related to: Didry et al. *The EMBO Journal* 31.4 (2012)



Complexes	Radius of gyration	Maximum diameter
	75.5 Å	285 Å
	55.5 Å	210 Å
	38.9 Å	130 Å
	25 Å	75 Å
	23.1 Å	70 Å

Histogram of intramolecular distances and ab initio molecular enveloppes determined using DAMMIF



Distance Distribution Function

Both the radius of gyration and the intensity at q=0 can be derived from p(r)

$$R_g^2 = \frac{\int_0^{D\max} r^2 p(r) dr}{2\int_0^{D\max} p(r) dr}$$

$$I(0) = 4\pi {r_e}^2 \varphi \int_0^D p(r) dr$$

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.

2017 publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution: an update

Acta Cryst. (2017). D73, 710–728

Jill Trewhella,[a]* Anthony P. Duff,[b] Dominique Durand,[c] Frank Gabel,[d] J. Mitchell Guss,[a] Wayne A. Hendrickson,[e] Greg L. Hura,[f] David A. Jacques,[g] Nigel M. Kirby,[h] Ann H. Kwan,[a] Javier Pérez,[i] Lois Pollack,[j] Timothy M. Ryan,[h] Andrej Sali,[k] Dina Schneidman-Duhovny,[l] Torsten Schwede,[m] Dmitri I. Svergun,[n] Masaaki Sugiyama,[o] John A. Tainer,[p] Patrice Vachette,[c] John Westbrook[q] and Andrew E. Whitten[b]

(*d*) Structural parameters.

	GI (tetramer)	BSA	CaM
Guinier analysis			
$I(0)$ (cm^{-1})	0.0759 ± 0.0008	0.0861 ± 0.0008	0.0554 ± 0.00008
R_g (Å)	32.87 ± 0.13	28.33 ± 0.05	21.74 ± 0.06
q_{min} (Å^{-1})	0.007	0.007	0.007
qR_g max (q_{min} = 0.0066 Å^{-1})	1.3	1.3	1.3
Coefficient of correlation, R^2	0.999	0.999	0.999
M from $I(0)$ (ratio to predicted)	178312 (1.03)	65589 (0.99)	21944 (1.31)
$P(r)$ analysis			
$I(0)$ (cm^{-1})	0.0748 ± 0.00008	0.0850 ± 0.00006	0.0533 ± 0.00006
R_g (Å)	32.65 ± 0.04	28.32 ± 0.03	22.2 ± 0.06
d_{max} (Å)	92	87	72
q range (Å^{-1})	0.007–0.243	0.007–0.282	0.0074–0.310
χ^2 (total estimate from *GNOM*)	0.929 (0.94)	0.858 (0.96)	0.855 (0.91)
M from $I(0)$ (ratio to predicted value)	180191 (1.04)	65354 (1.00)	21718 (1.29)
Porod volume (Å^{-3}) (ratio V_P/calculated M)	229000 (1.3)	101000 (1.5)	25200 (1.5)
V, M using the Fischer method (ratio of M to expected)	192400, 157.9 (0.91)	82440, 67.9 (1.02)	21550, 17.7 (1.05)

A FEW EXPERIMENTAL CONSIDERATIONS

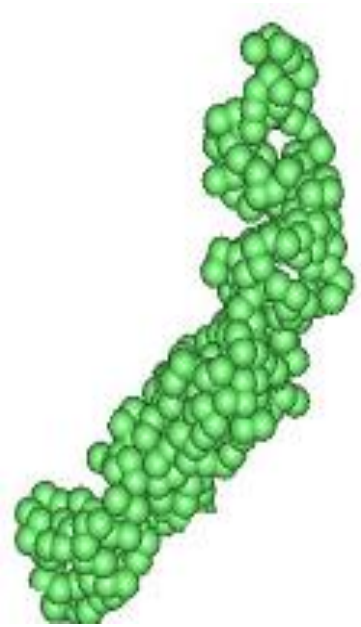
Evaluation of the solution properties

$i_1(q)$

Ideality

$I(q)$

Monodispersity



One must check that both assumptions are valid for the sample under study.

molecule

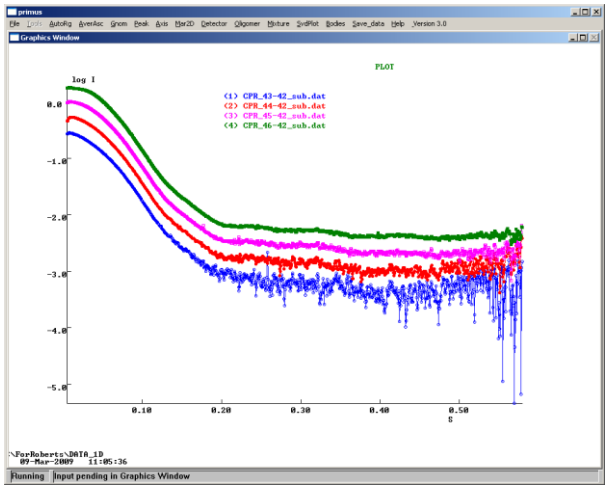
experimental

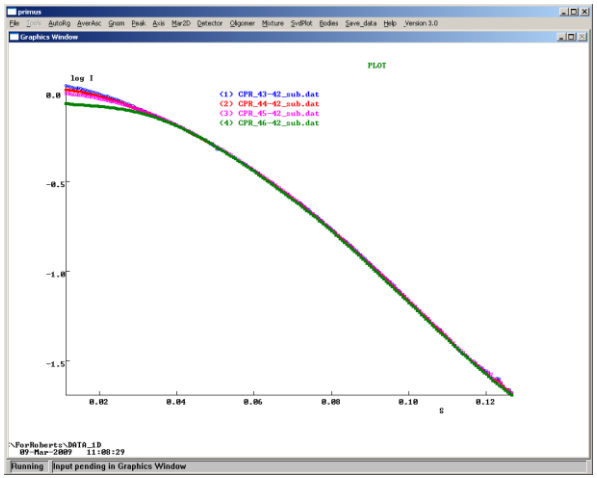
Evaluation of the solution properties

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 check whether the scattering pattern is independent of concentration.





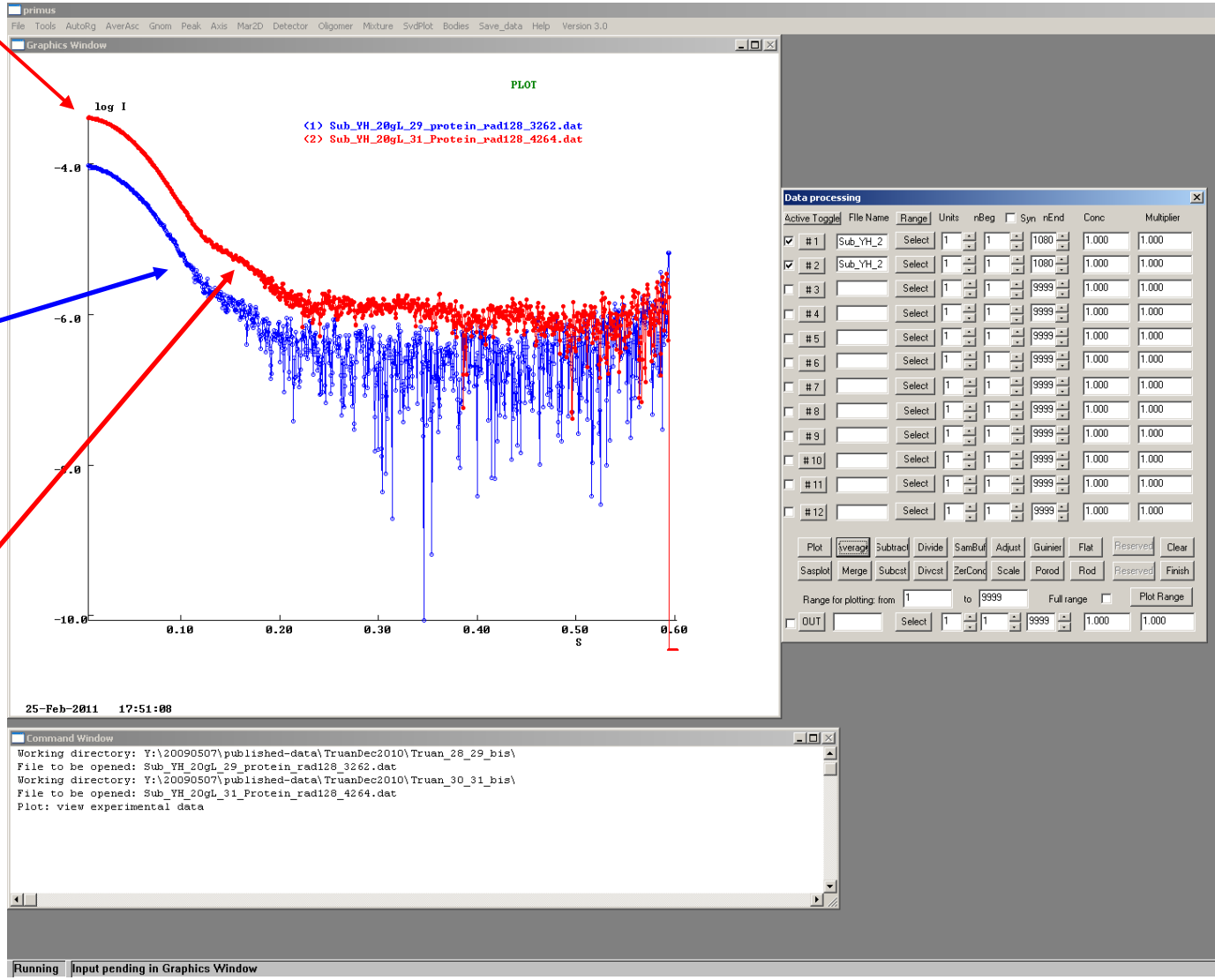
Merging data with “Primus” from ATSAS

Merge low c (@ low q) and high c (@ high q) curves

Here, slight repulsive interactions alter the high concentrated solution curve at small angles

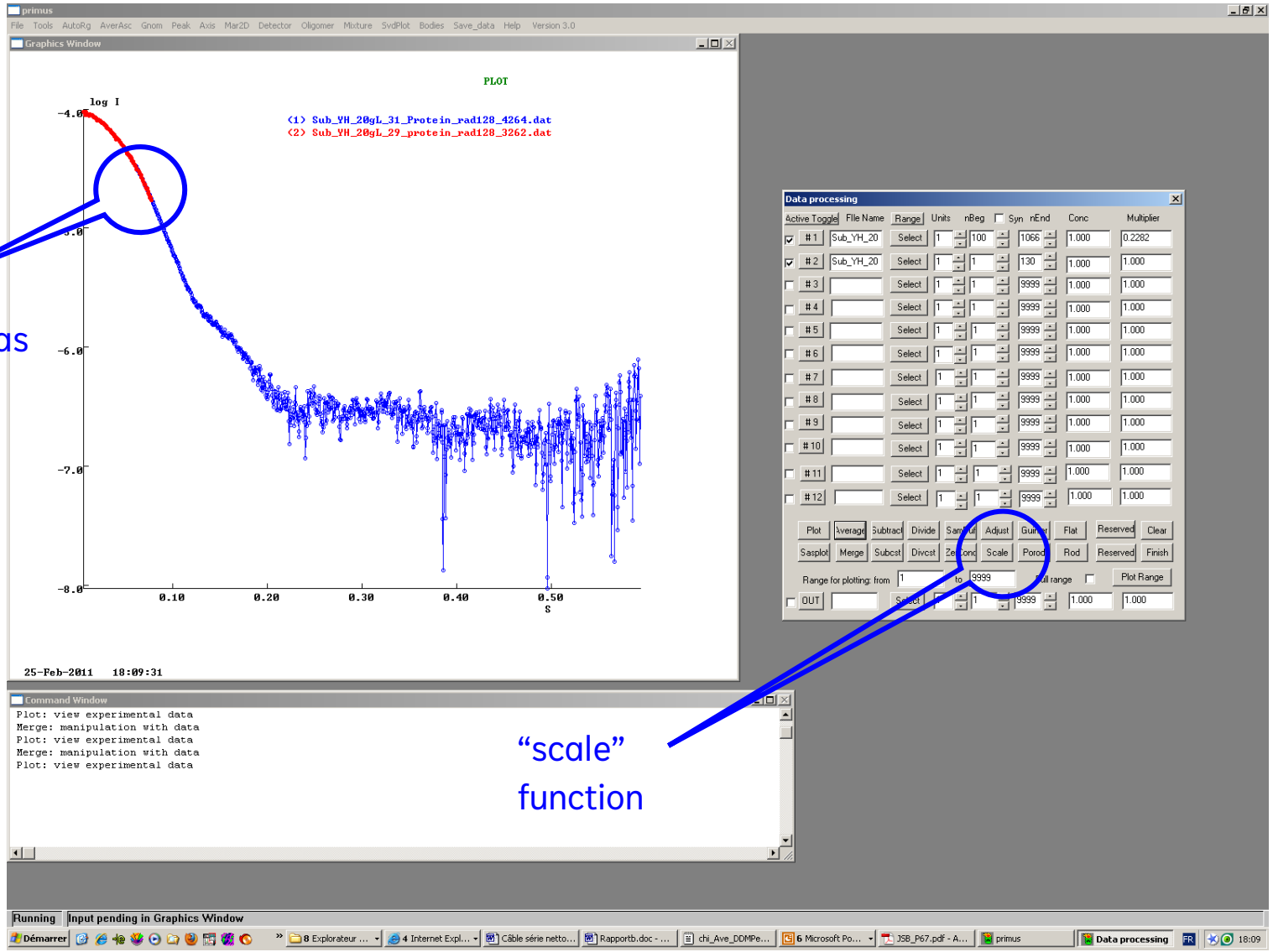
Small angle data using the lowest concentration curve or an extrapolation to zero concentration from a series of dilute solutions (correction of interparticle effects)

larger angle data using the most concentrated solution



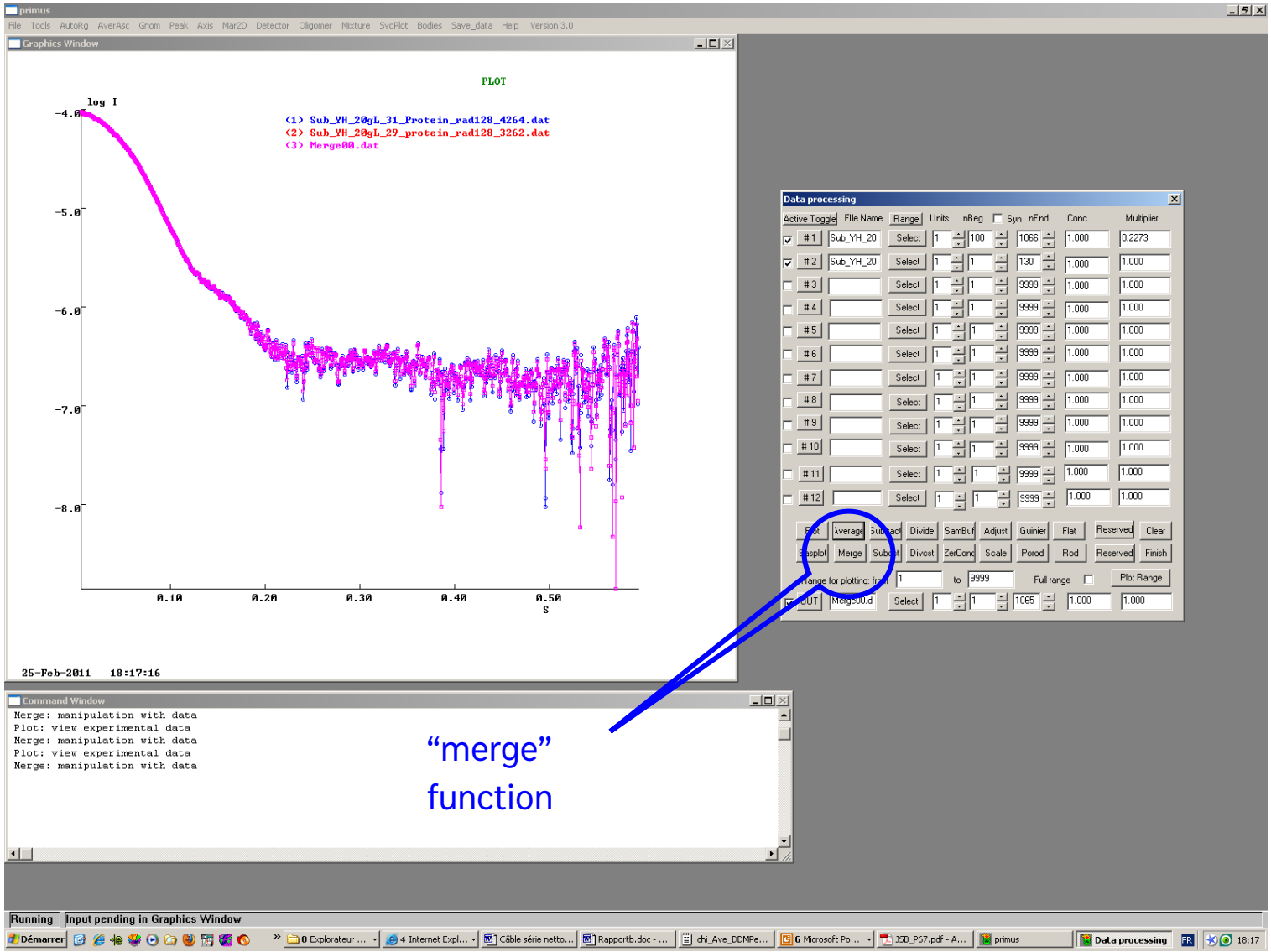
Merging data with “Primus” from ATSAS

The common range should be as restricted as possible to avoid adding noise



“scale” function

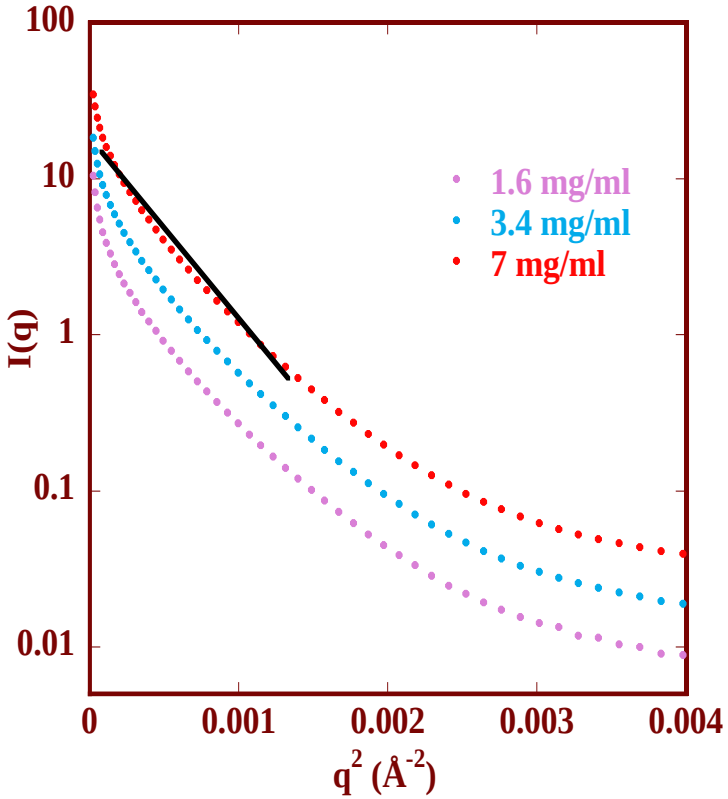
Merging data with “Primus” from ATSAS



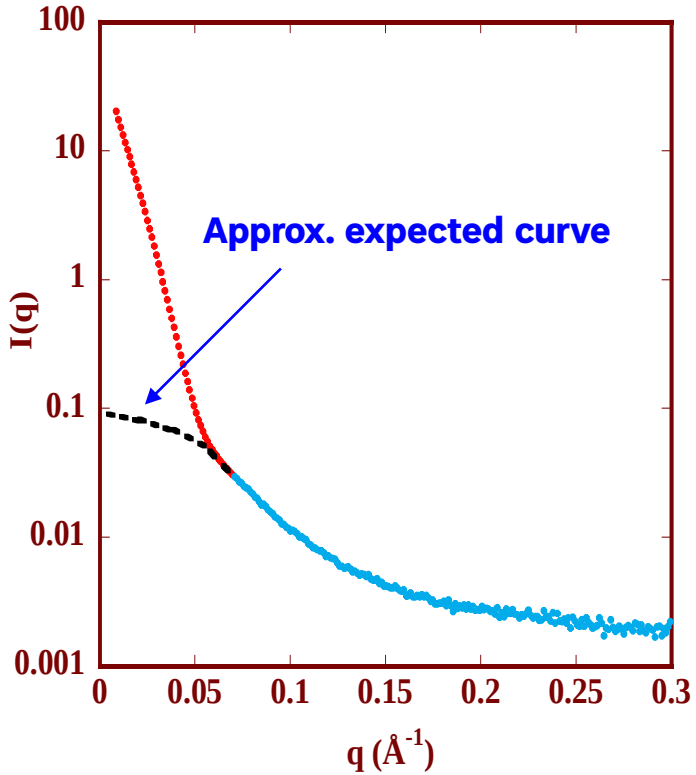
Evaluation of the solution properties

Irreversible aggregation

→ Useless data: the whole curve is affected



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



Swing – Domaine 1-242 de RRP44 – 07/08

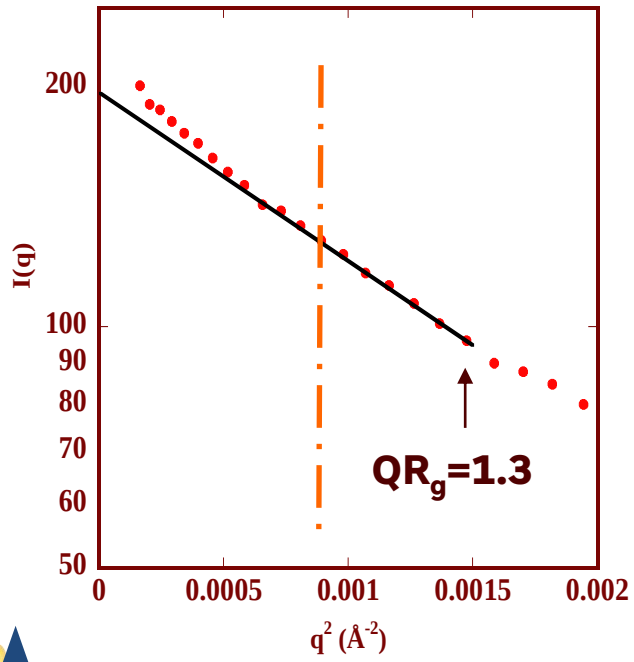
(Courtesy D. Durand, IBBMC, Orsay)

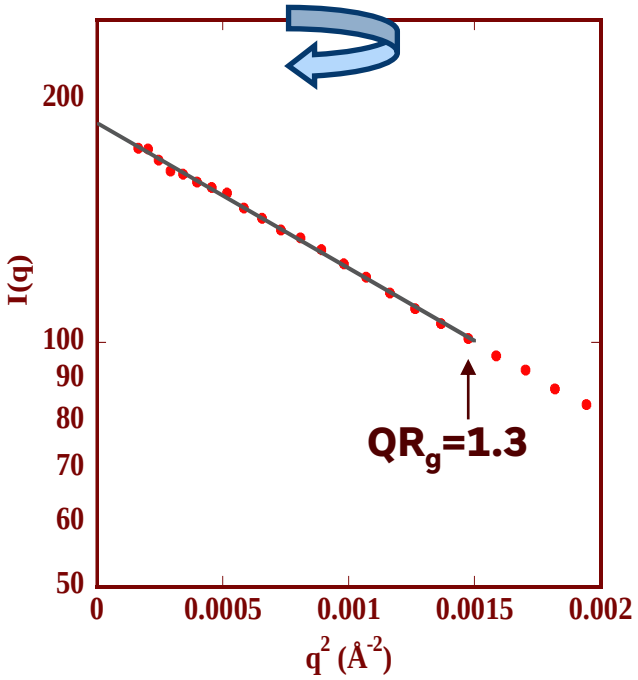
Evaluation of the solution properties

Weak aggregation → **possible improvement**

centrifugation, buffer change

Nanostar –PR65 protein





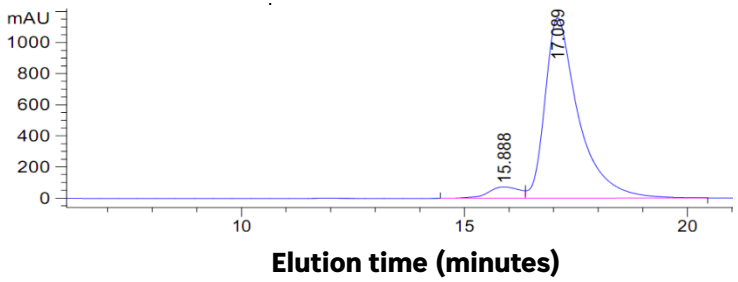
R_g ~ 38 Å – slightly too high

R_g ~ 36 Å

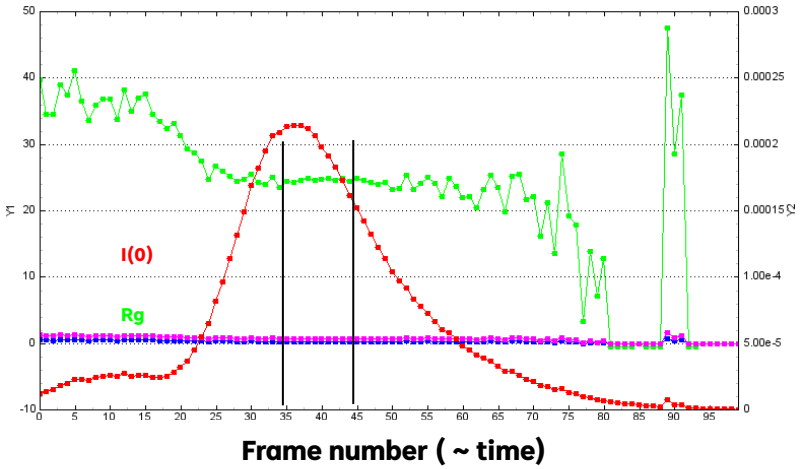
(Courtesy D. Durand, IBBMC, Orsay)

Remove aggregates using SEC-SAXS

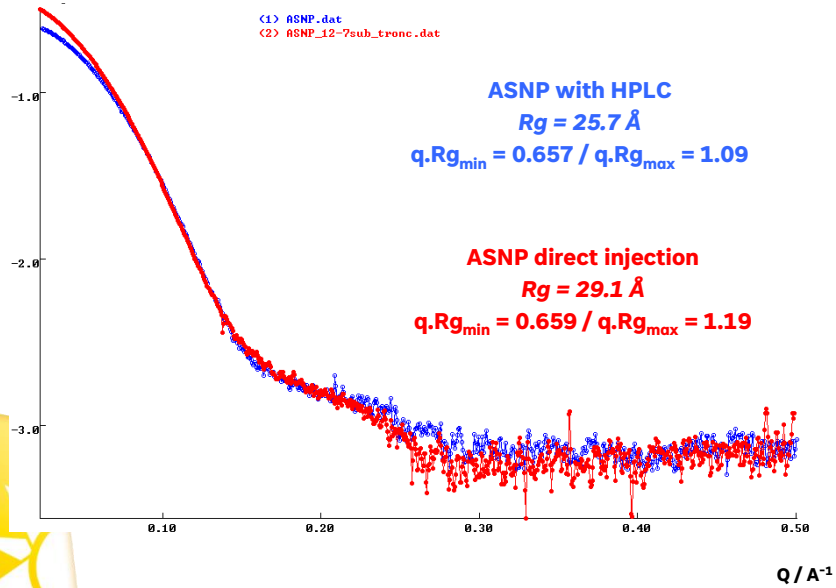
ASNP elution profile, monitored by UV absoprtion at 280 nm



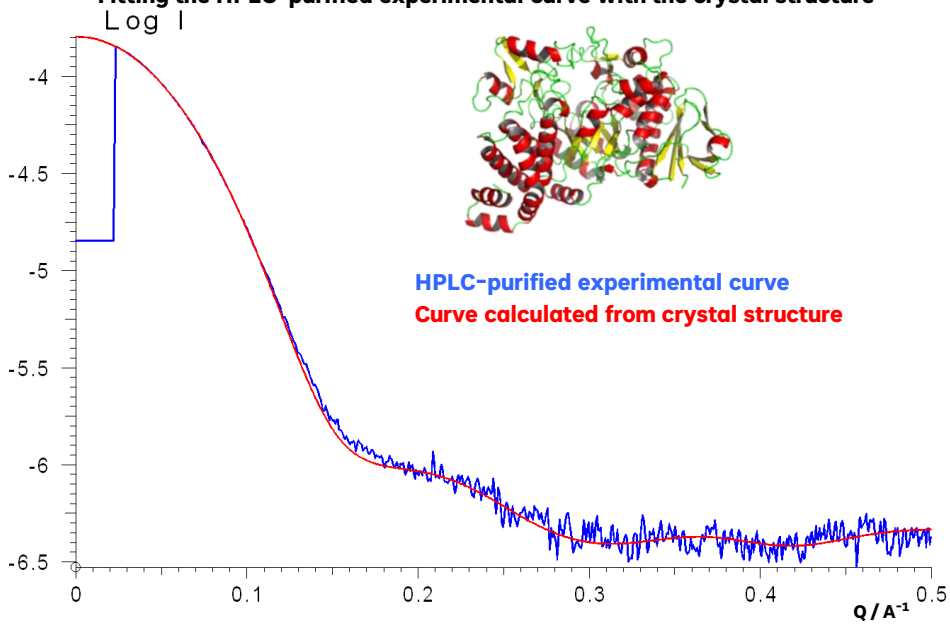
I(0) and Rg determined for each SAXS frame during elution



Comparison between HPLC-purified and Direct injection curves

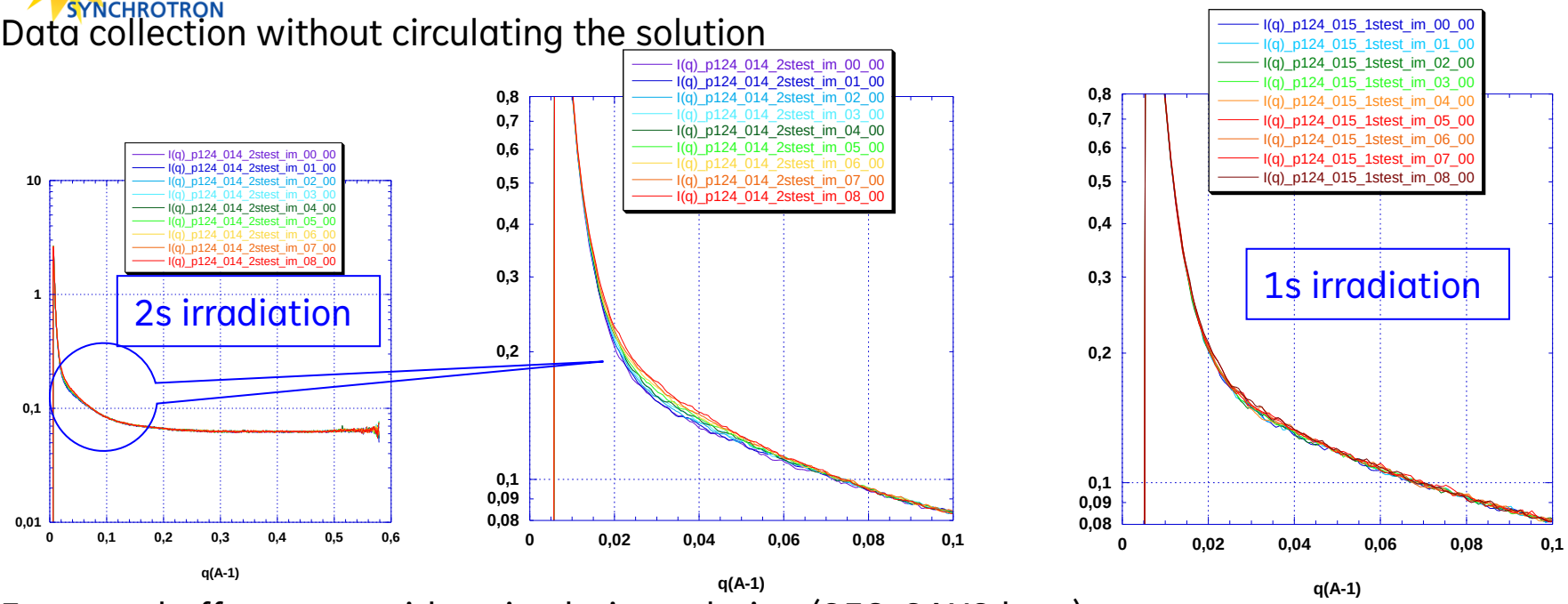


Fitting the HPLC-purified experimental curve with the crystal structure

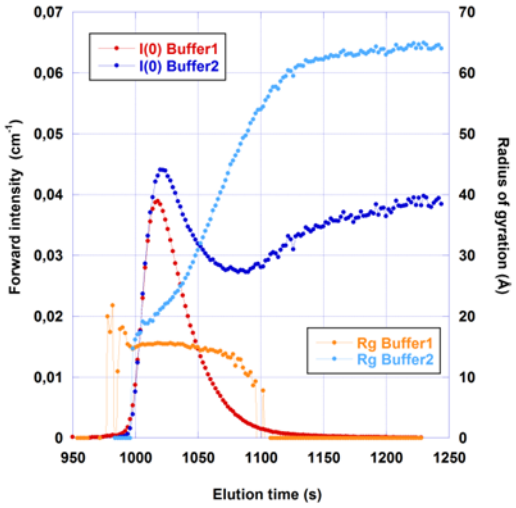


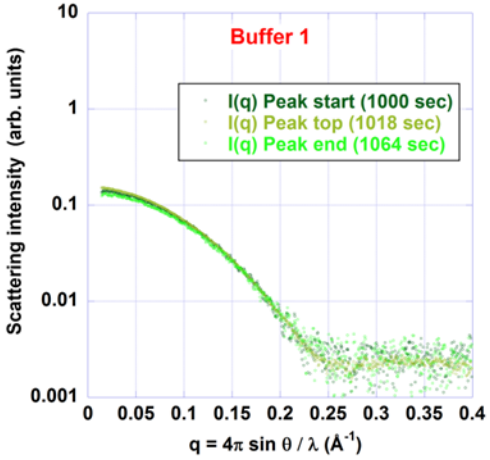
Unwanted Radiation effects

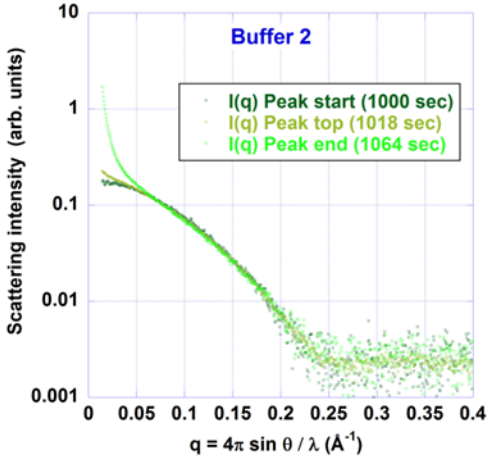
Data collection without circulating the solution



For some buffers, even with a circulating solution (SEC-SAXS here)



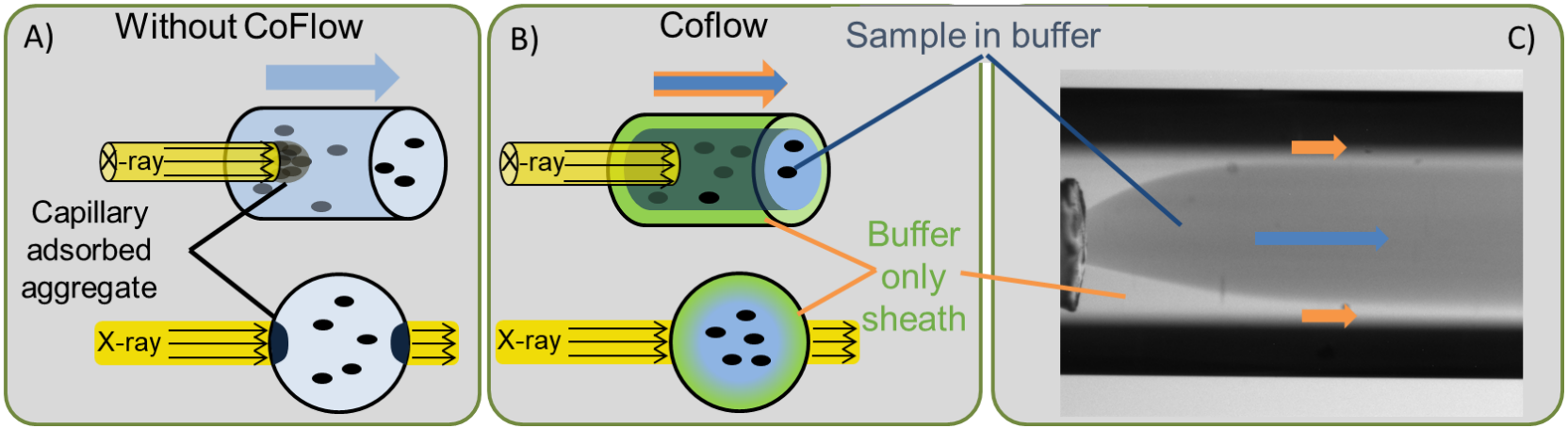




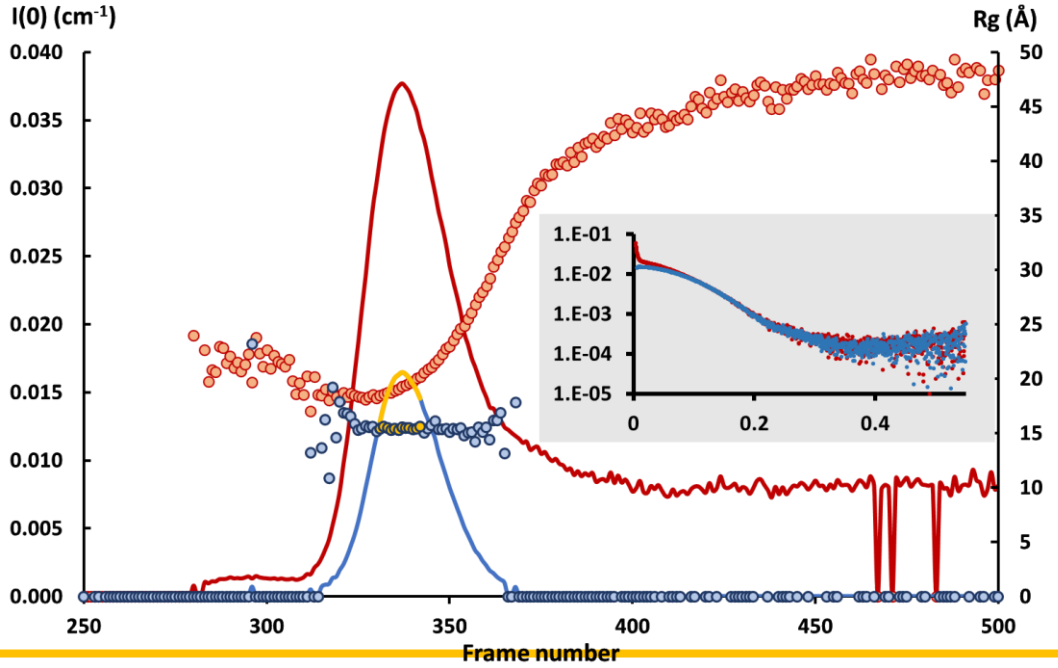
Pérez J., Vachette P. (2017) In: Biological Small Angle Scattering: Techniques, Strategies and Tips. Advances in Experimental Medicine and Biology, vol 1009. Springer, Singapore

Unwanted Radiation effects

Co-flow set-up

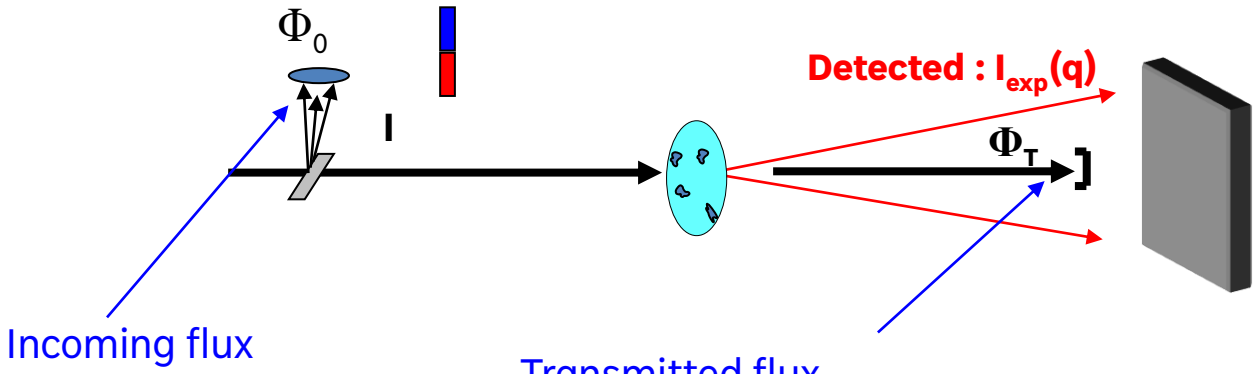


Kirby, N *et al.* (2016). Acta Cryst. D72, 1254–1266.

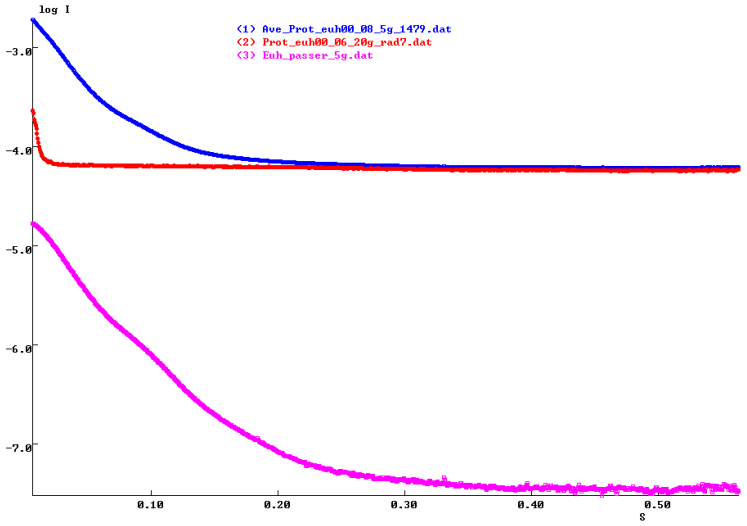


Transmission and buffer measurements are crucial

- 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, ...).



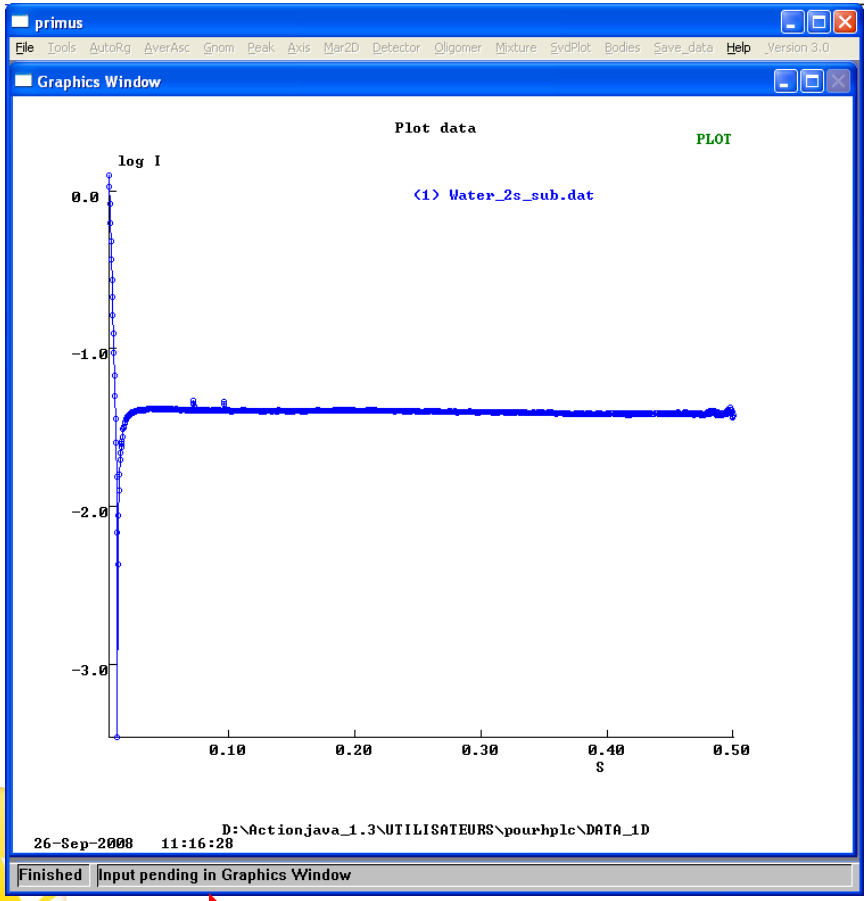
$$I_{particles}(q) = I_{sample}(q) - I_{buffer}(q)$$

Calibration of the set-up using water scattering

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

Molecular density — Isothermic compressibility

$I_{H2O,theory}$ = 0.0163 cm^{-1} at 20°C



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

Example:

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

$I_{H2O,exp}$ = 0.042 Exp. Units

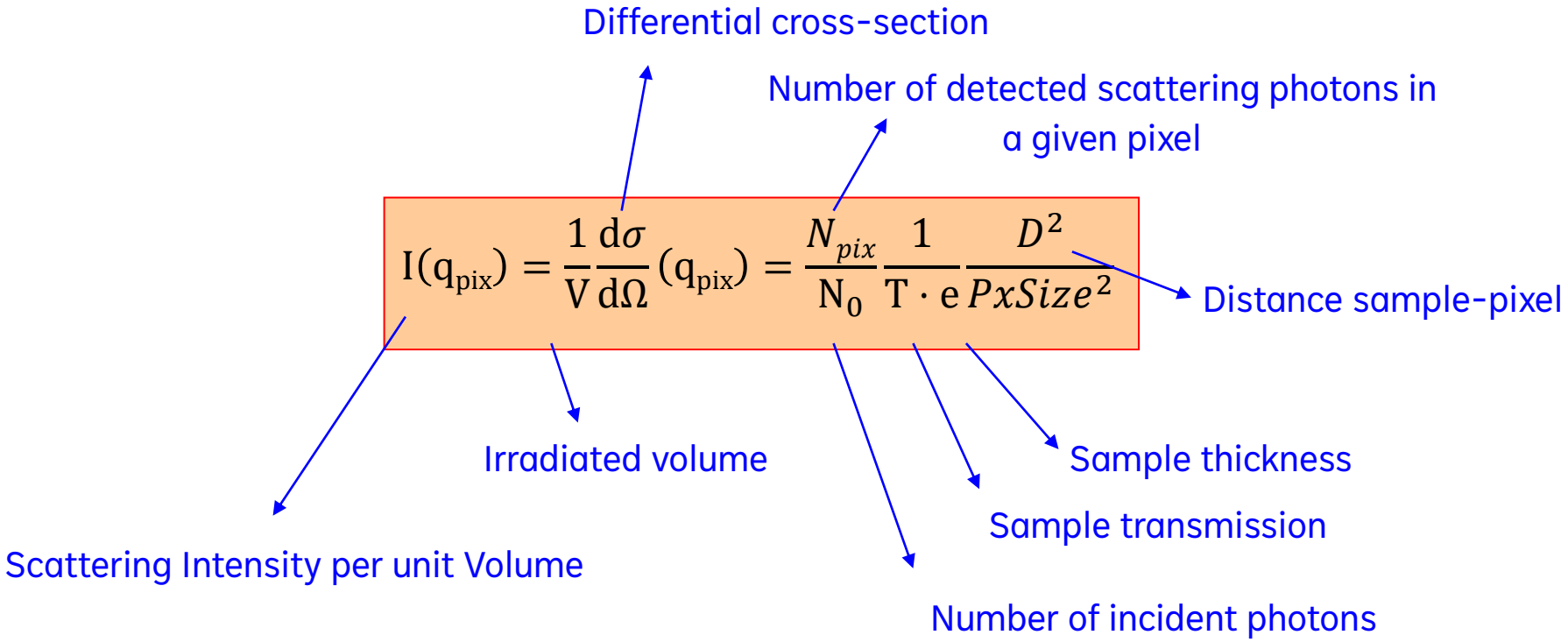
$I_{H2O,exp} = K_{exp} * I_{H2O,theory}$

→ Here : K_{exp}=2.56 Exp.Units / cm^{-1}

For any sample in that capillary : $I_{theory}(cm^{-1}) = I_{exp} / K_{exp} = I_{exp} / 2.56$

From measurements to I(q) data

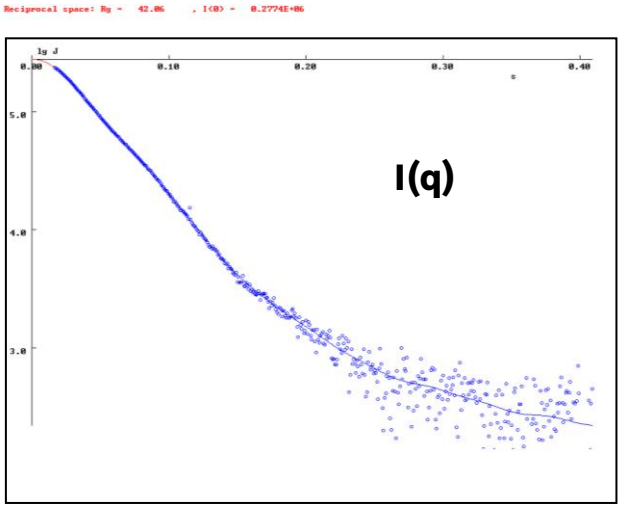
Relation between the the number of measured photons N_{pix} on a given pixel of the detector, making a solid angle $\Delta\Omega$, and the Scattering Intensity per unit volume :

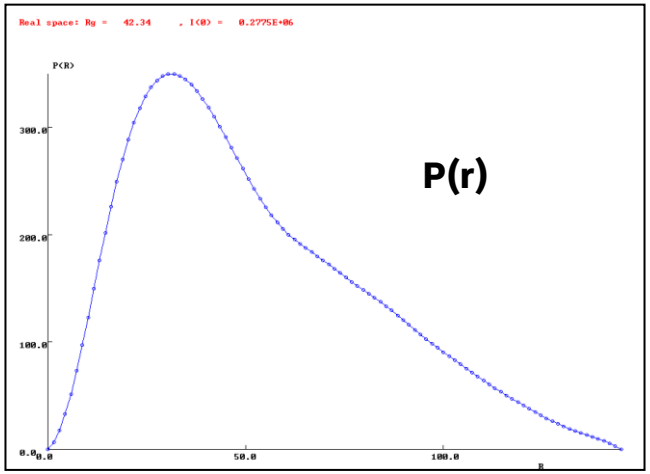


At this stage

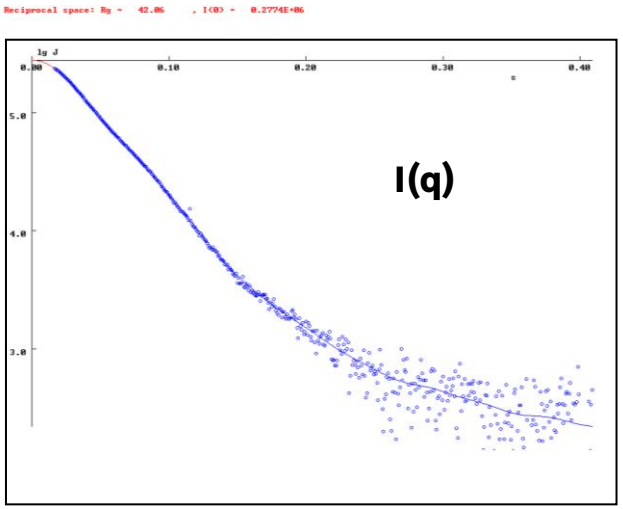
We have gone from

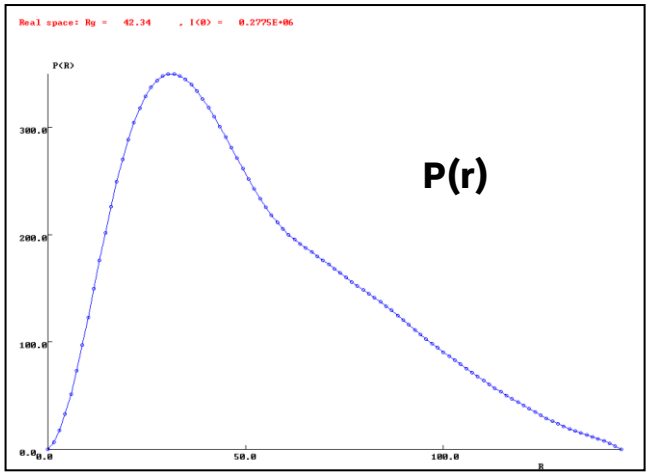
to



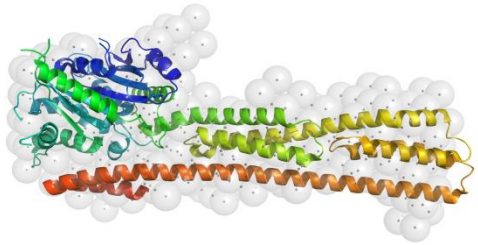


Now, we have to go from





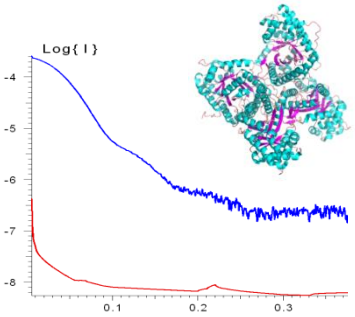
to



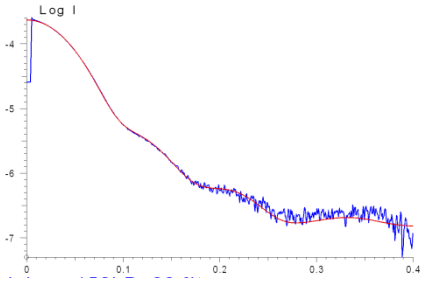
MODELLING

BioSAXS data modeling

- Theorical model or complete atomic structure available

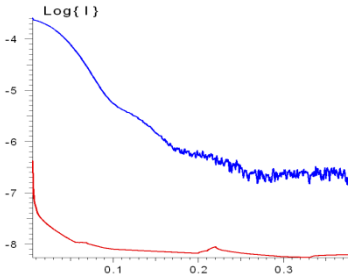


Validation/identification in solution

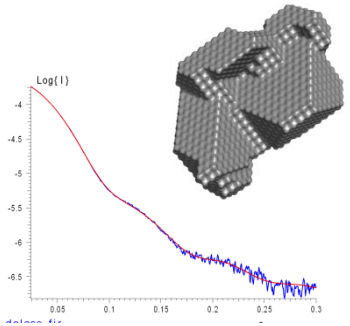


CRYSOL
FOXS
PepsiSAXS
WAXSiS

- Nothing known (except the curve)

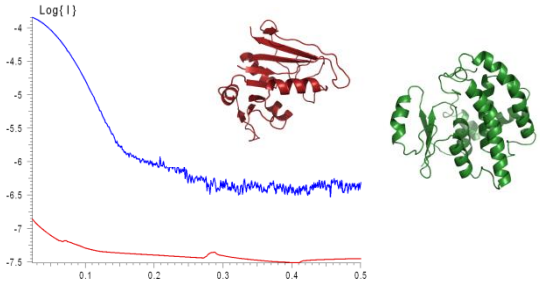


Low resolution model



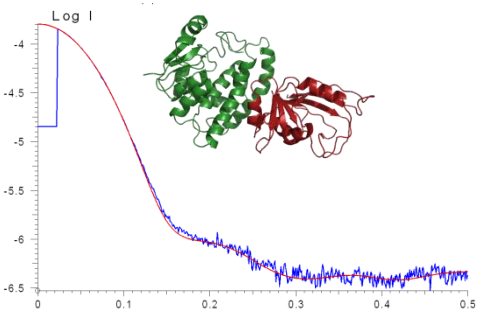
DAMMIN
DAMMIF
GASBOR
MONSA
DENFERT

- Structures of subunits available



Rigid body modeling of the complex and

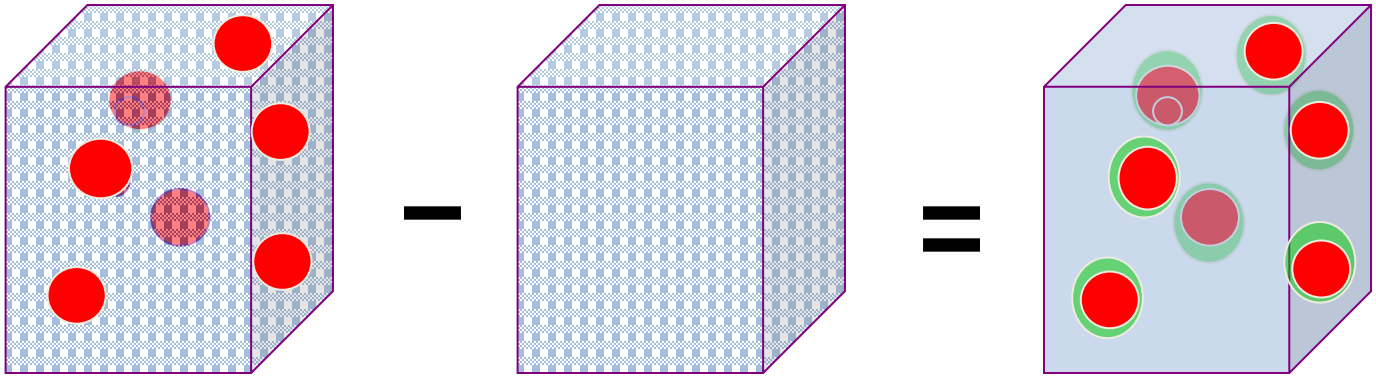
molecular modeling of the missing part



SASREF
BUNCH
CORAL
DADIMODO

Slide slightly modified from Dmitri Svergun, EMBL Hamburg

Solvent scattering and contrast



$I_{solution}(q)$ $I_{solvent}(q)$ $I_{particle}(q)$

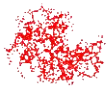
To obtain scattering from the particles, solvent scattering must be subtracted to yield the pair distribution of the effective density $\Delta\rho(r) = \rho(r) - \rho_0$, where ρ_0 is the scattering density of the solvent.

Further, the bound solvent density may differ from that of the bulk.

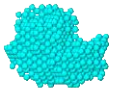
CRYSOL: from atomic coordinates to a SAXS curve

$$I_{calc}(q) = \langle |A_a(\vec{q}) - \rho_S A_s(\vec{q}) + \delta\rho_b A_b(\vec{q})|^2 \rangle_\Omega$$

$A_a(\boldsymbol{q})$ = *molecular scattering amplitude in vacuum*



$A_s(\boldsymbol{q})$ = *scattering amplitude from excluded volume*



$A_b(\boldsymbol{q})$ = *scattering amplitude from the hydratation 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 :

$$I_{calc}(q) = \sum_{l=0}^{L} \sum_{m=-1}^{l} \left| A_{lm}(q) - \frac{V}{V_{calc}} \rho_S C_{lm}(q) + \delta\rho B_{lm}(q) \right|^2$$

The experimental scattering curves are then fitted using only 3 parameters in order to minimize the discrepancy χ :

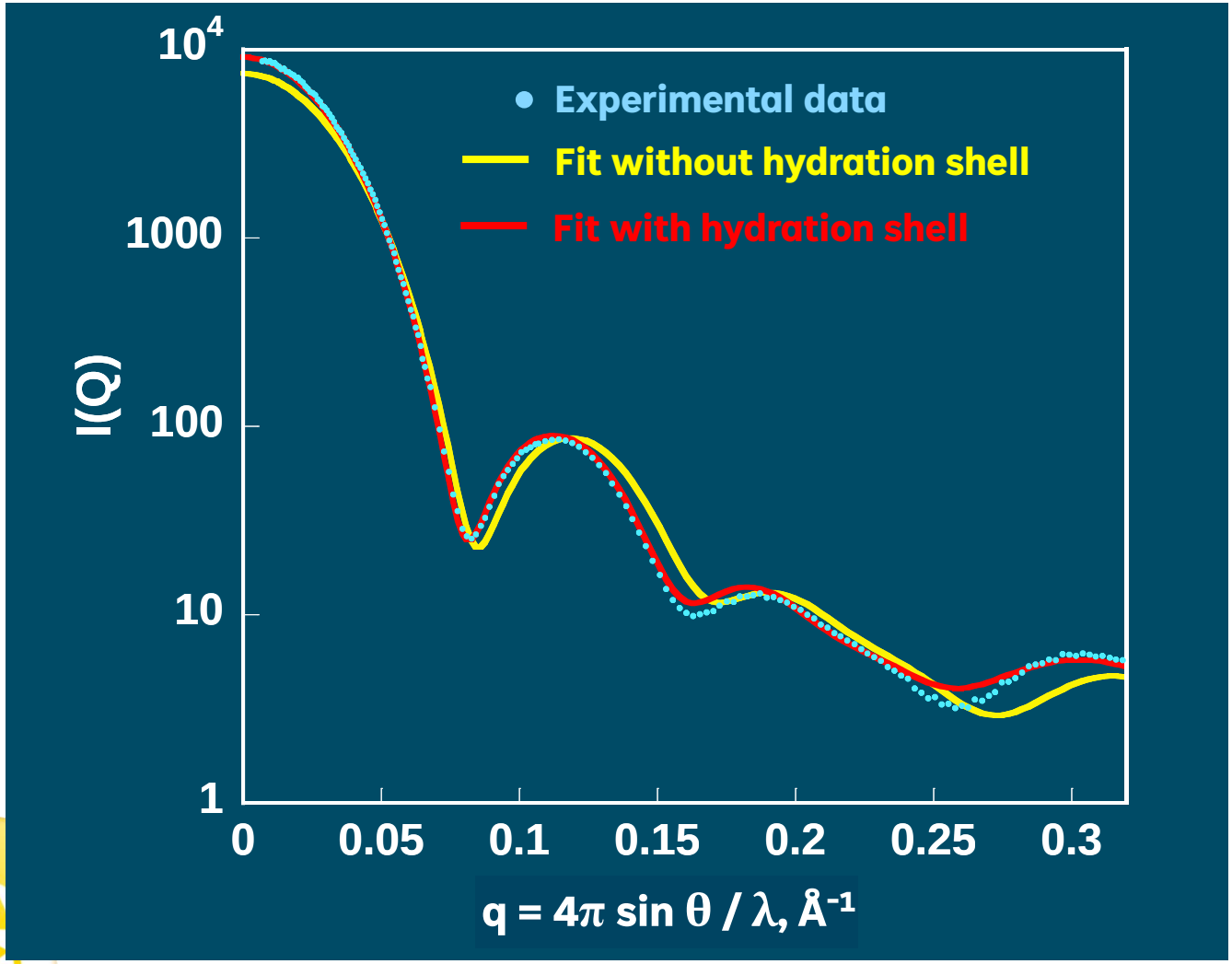
- the general scale of $I_{calc}(q)$
- the total excluded volume V, which is equivalent to modifying the average electronic contrast
- the contrast of the border layer $\delta\rho$

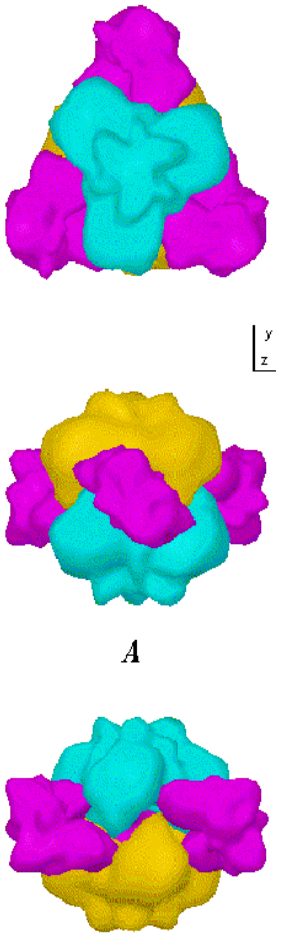
$$\chi^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left[\frac{I_{\exp}(q_i) - scale * I_{calc}(q_i)}{\sigma_{\exp}(q_i)} \right]$$

Svergun , Barberato & koch (1995), J. Appl. Cryst., ***28****, 768*

Effect of the hydration shell

T state of *E. coli* allosteric ATCase





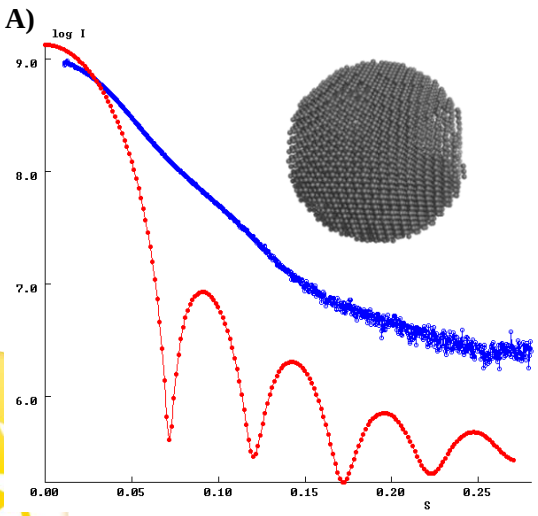
3D shape reconstructions from SAXS data with DAMMIN

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 radius (r_b)

Starting model = sphere with a radius R = $D_{max}/2$ with N scattered beads ($r_b << R$)
The number of the "dummy atom" $N \approx (R/r_b)^3$

Each sphere 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)

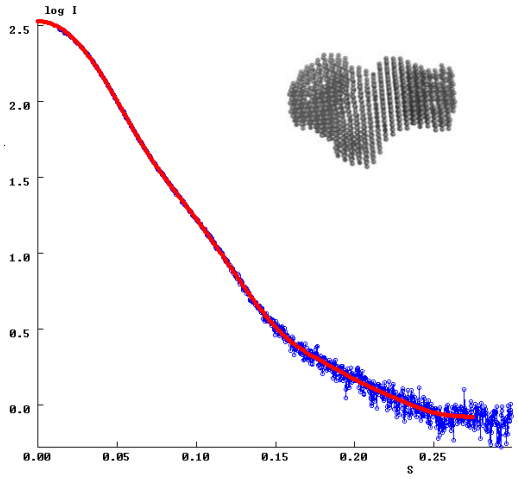


$$f(X) = \chi^2\left[I(q)_{\exp}, I(q,X)\right] + \alpha P(X)$$

X is a conformation of the system
P(X) is a penalty function

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

After k iterations



D. I. Svergun, M. Kozin, M. Petoukhov, V. Volkov (1999). Biophys J. 2879-2886.

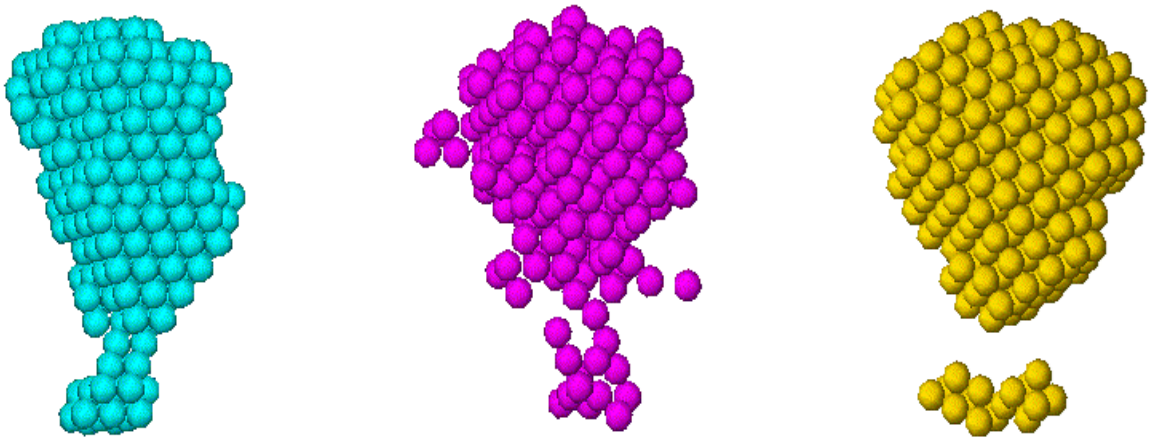
3D shape reconstructions from SAXS data with DAMMIN

- Obtaining 3D shapes from SAXS data is an 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

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

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



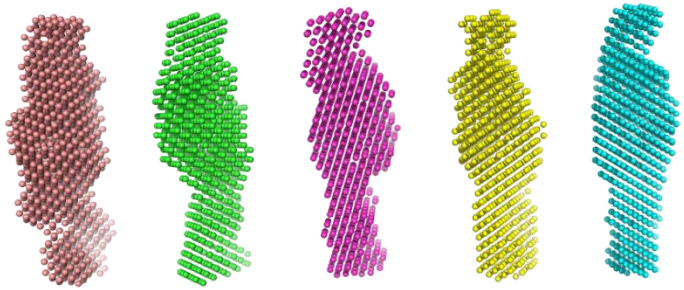
compact

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



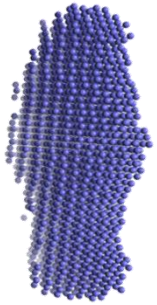
Shp1 Shp2 Shp3 Shp4 Shp5

File	Aver	1	2	3	4	5
1	0.52	0.00	0.51	0.52	0.50	0.52
2	0.52	0.51	0.00	0.52	0.49	0.52
3	0.53	0.52	0.52	0.00	0.53	0.52
4	0.53	0.50	0.49	0.53	0.00	0.54
5	0.53	0.52	0.52	0.52	0.54	0.00

Mean value of NSD : 0.535
Standard deviation of NSD : 0.008

Damsel.log

Damfilt
(intersection)

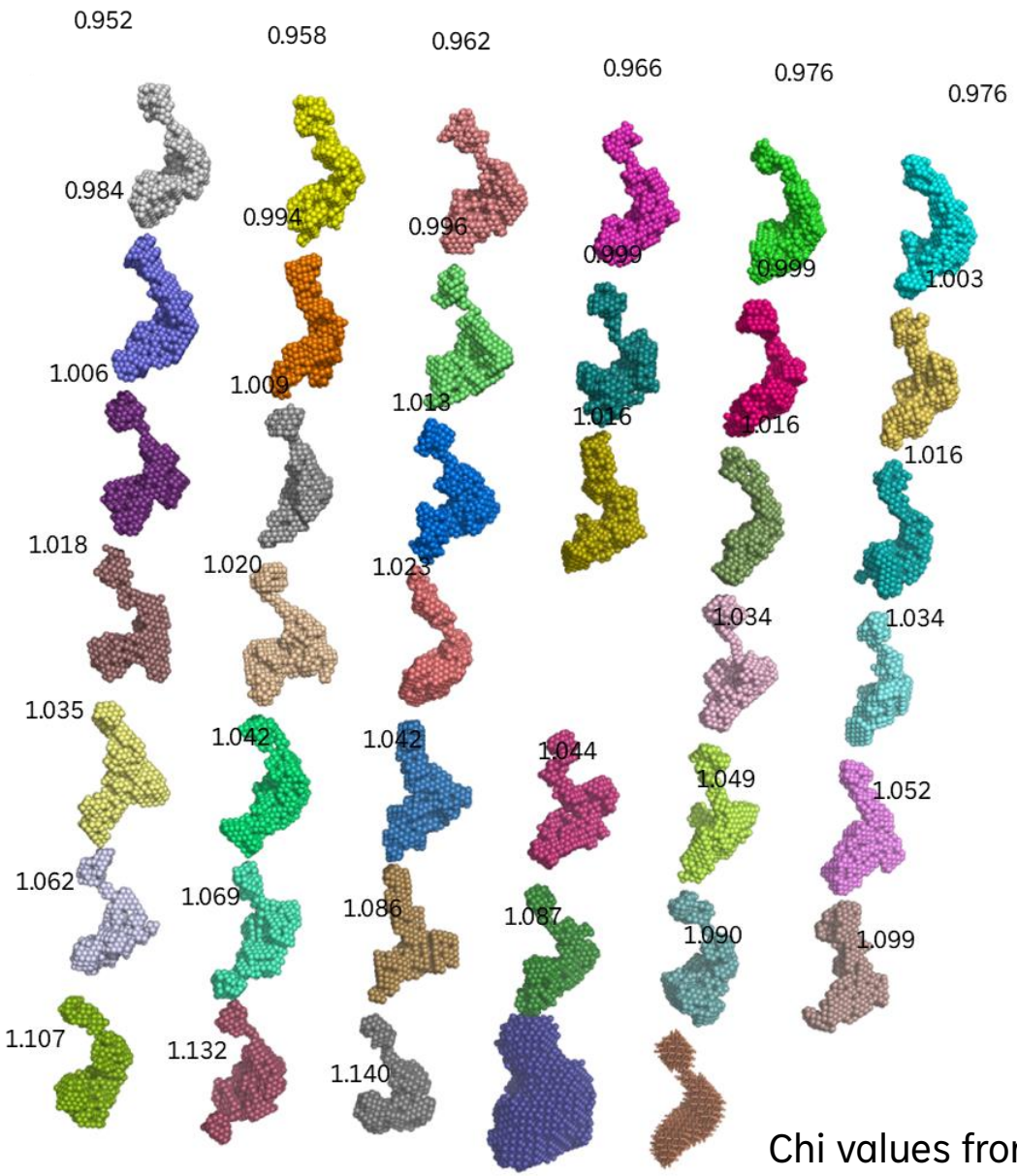


Damaver
(all superimposed)

- 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.pdb because $I_{damfilt}(q) \neq I_{exp}(q)$

3D shape reconstructions from SAXS data with DAMMIN

NSD values

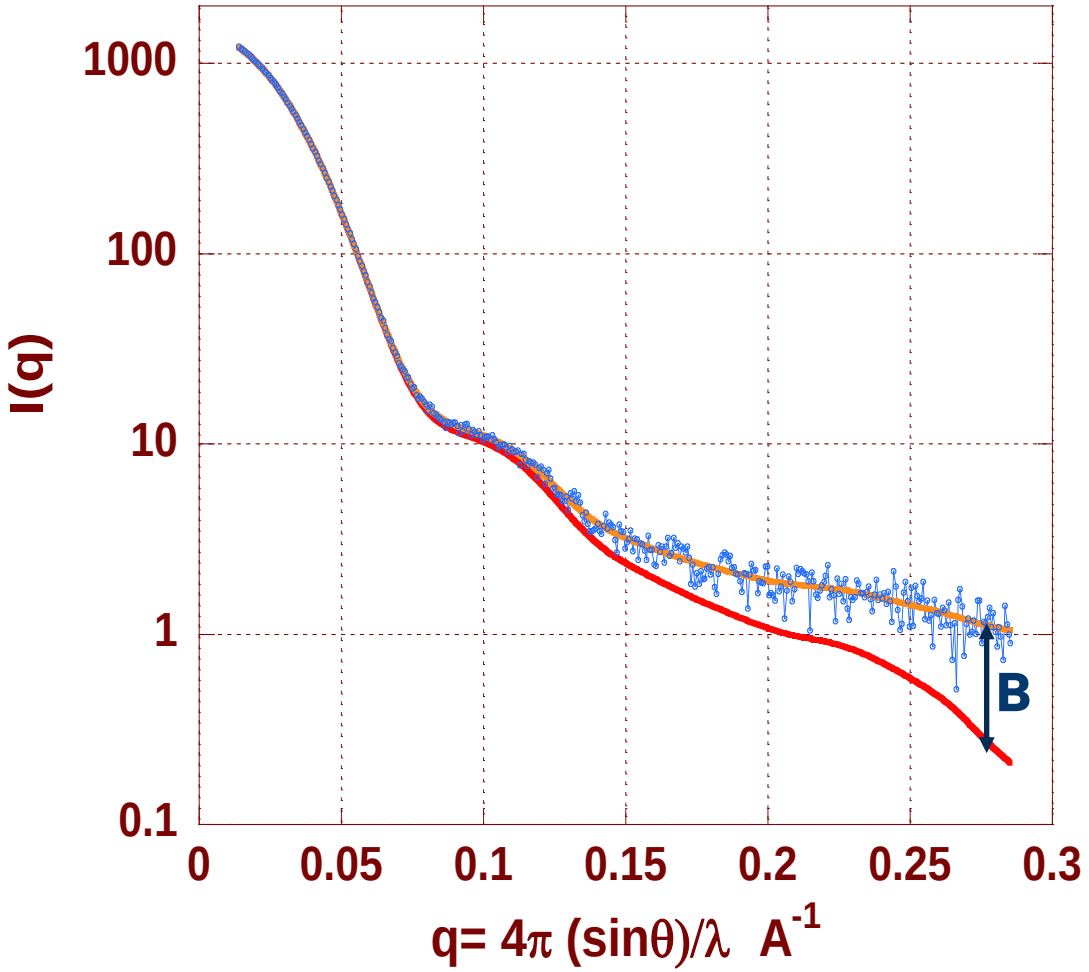


Chi values from 1.704 to 1.942

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.



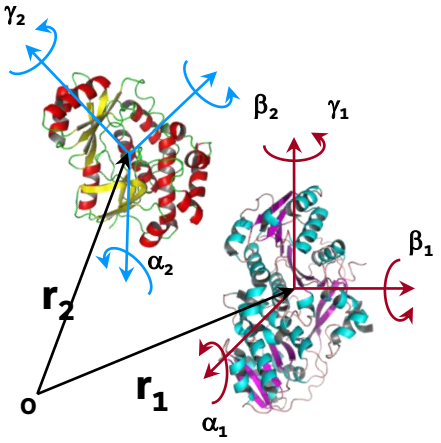
SASREF: Rigid body modeling against SAXS data

SASREF : when atomic structures of domains are known, but not their mutual organization

The objective is to find the relative orientation of each subunit with a correct 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 each subunit

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



The **amplitudes** are calculated with CRYSOL from the high resolution structure of each monomer

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.

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

+ CORAL, BUNCH: missing residues are modelled as beads

Petoukhov & Svergun (2005). Biophys. J., 89, 1237-1250.

DADIMODO : Data-Driven Modules Docking

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

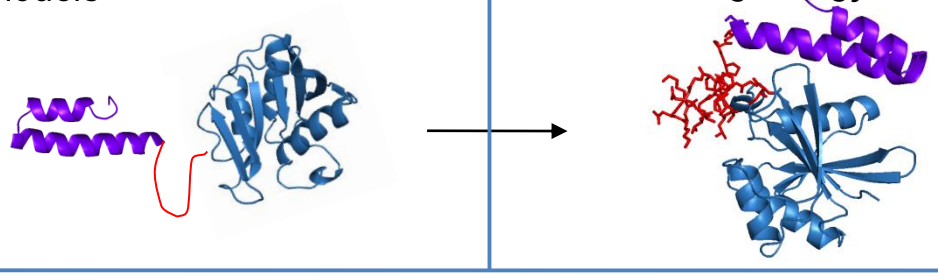
F. Mareuil, et al. (2007) *Eur Biophys J.*

Evrard et al. (2011), *J. Appl. Cryst.*

Modelling approach : complete atomic model

Full structure initiated with :

- Crystal or NMR domain structures
- Homology models



Prior knowledge:

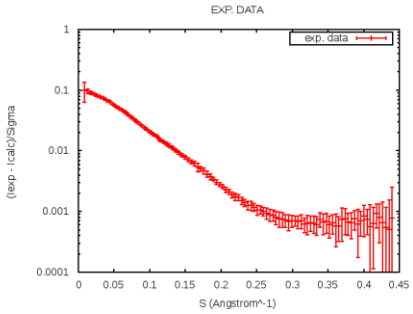
- Sequence
- Sub-parts moved as rigid-bodies (user-defined)
- A correct stereochemistry is maintained at all steps by minimizing energy

Experimental data:

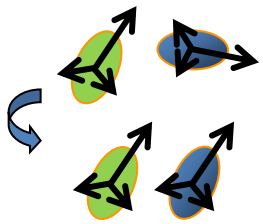
- SAXS
- NMR

  RDC

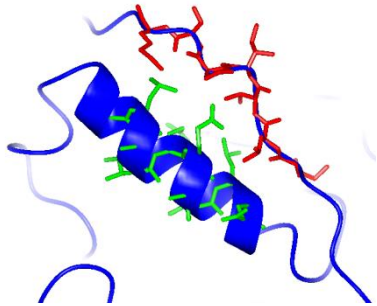
  ADR (chem. shift map.)



SAXS score



RDC score



ADR score

Optimisation of the **all-atom** structure *via* a genetic algorithm

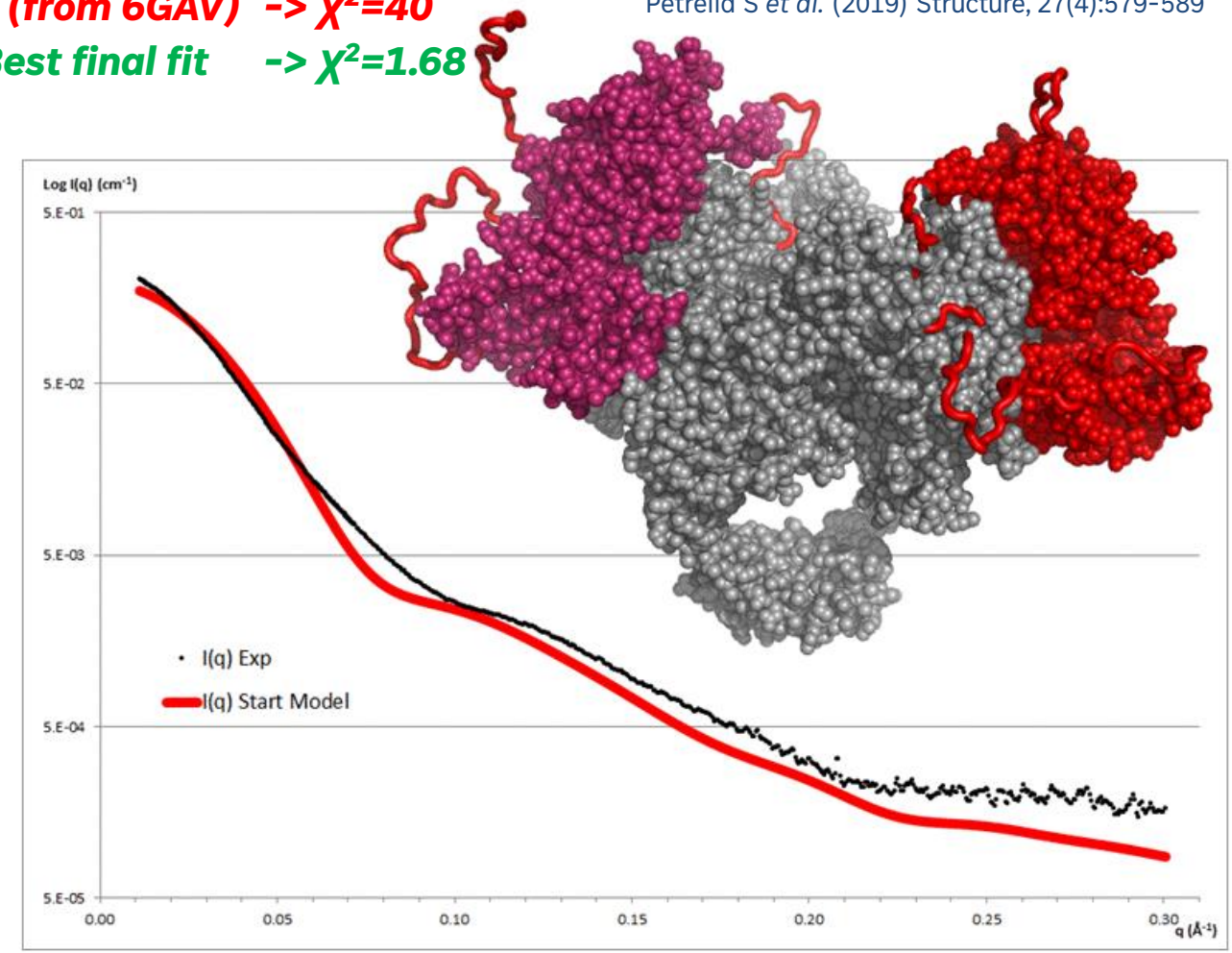
Example: DNA Gyrase

Mycobacterium tuberculosis DNA Gyrase

Petrella S *et al.* (2019) Structure, 27(4):579-589

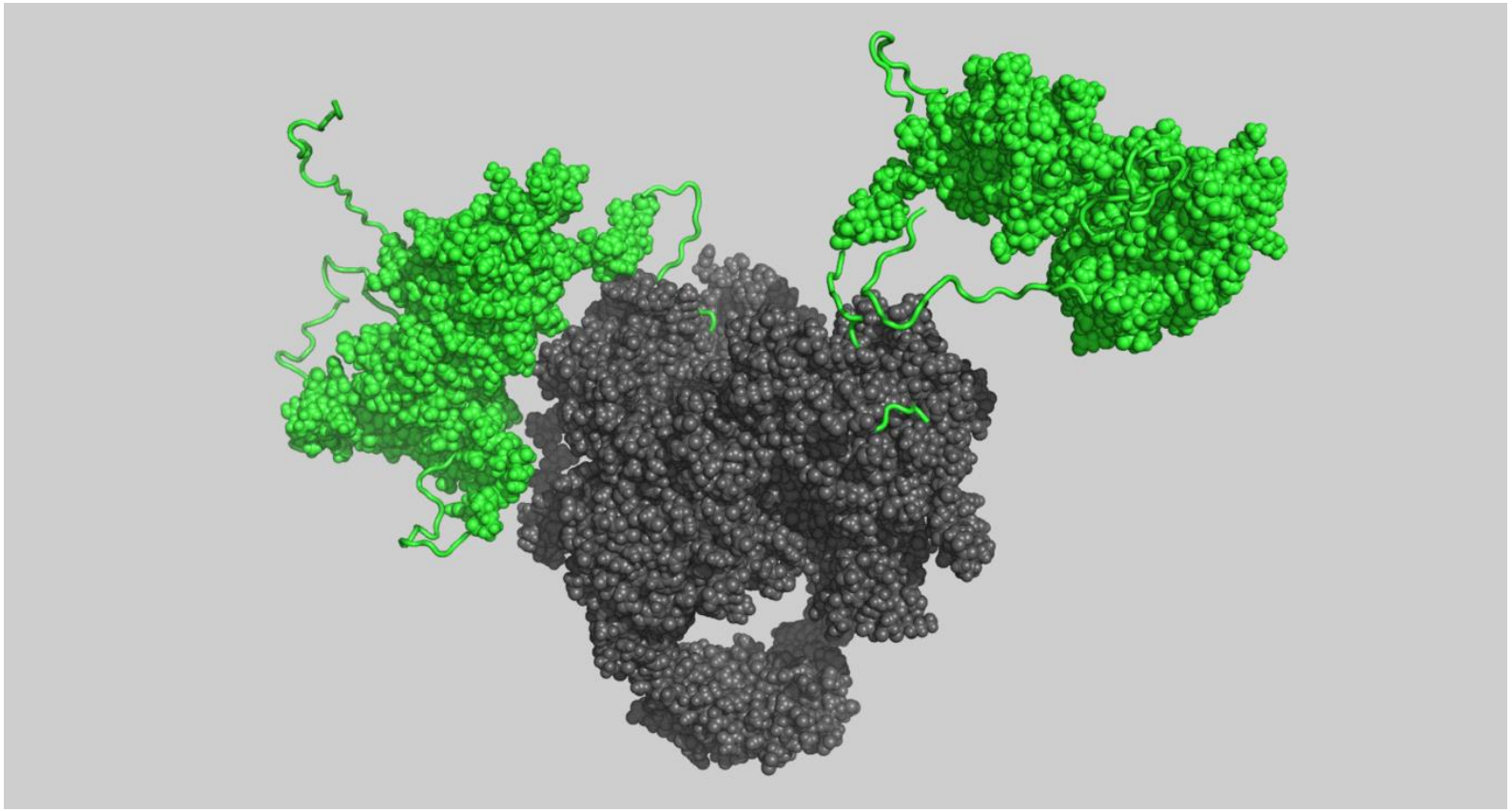
Start model (from 6GAV) -> χ^2=40

Best final fit -> χ^2=1.68



Example: DNA Gyrase

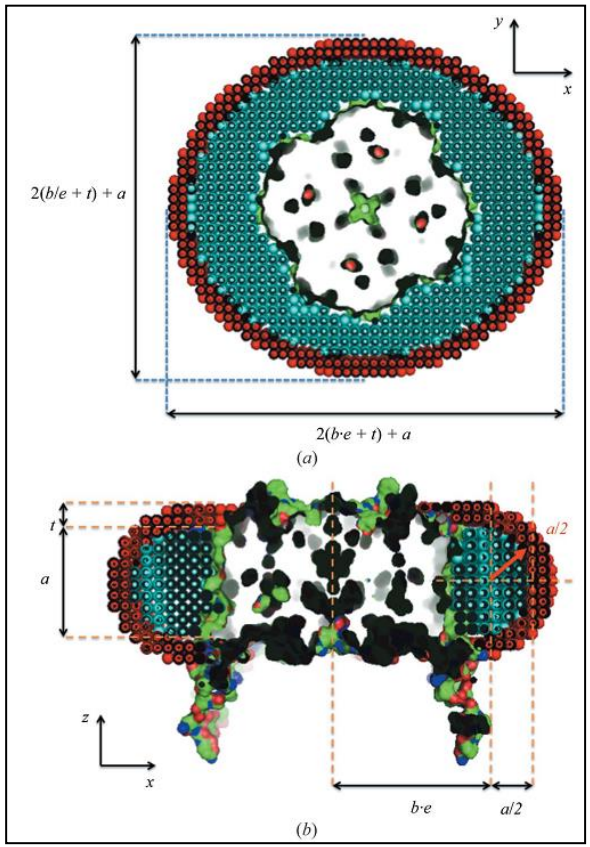
5 best final fits : 1.68 < χ^2 < 1.76



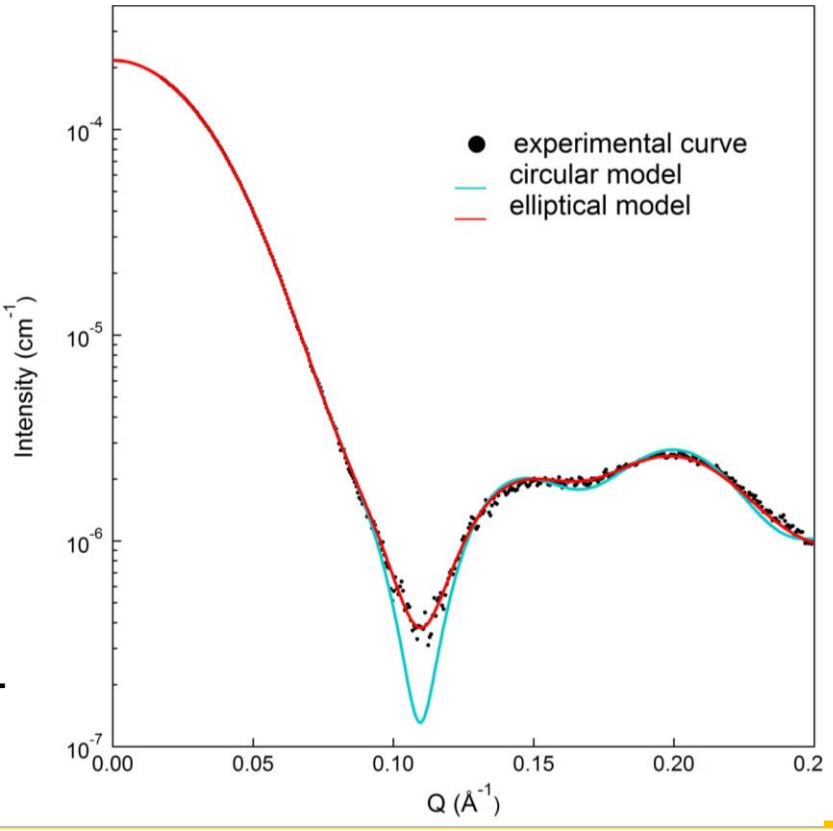
Mycobacterium tuberculosis DNA Gyrase

Memprot : a program to generate & optimize a detergent corona model

Pérez J. & Koutsioubas, A. (2014). Acta Cryst.D70
F. De Pol et al. (2024), J. Appl. Cryst 57



Algorithm of the *Memprot* program. The program essentially creates PDB files with the models made of the full-atom protein structure and the parameterized coarse-grained detergent corona, and *CRYSOL* is called to calculate the SAXS curves. An overall sorting on the χ value is performed to keep the best model.



Berthaud et al. (2012), *JACS*, 134 (24), 10080-10088

A FEW CONCLUDING MESSAGES

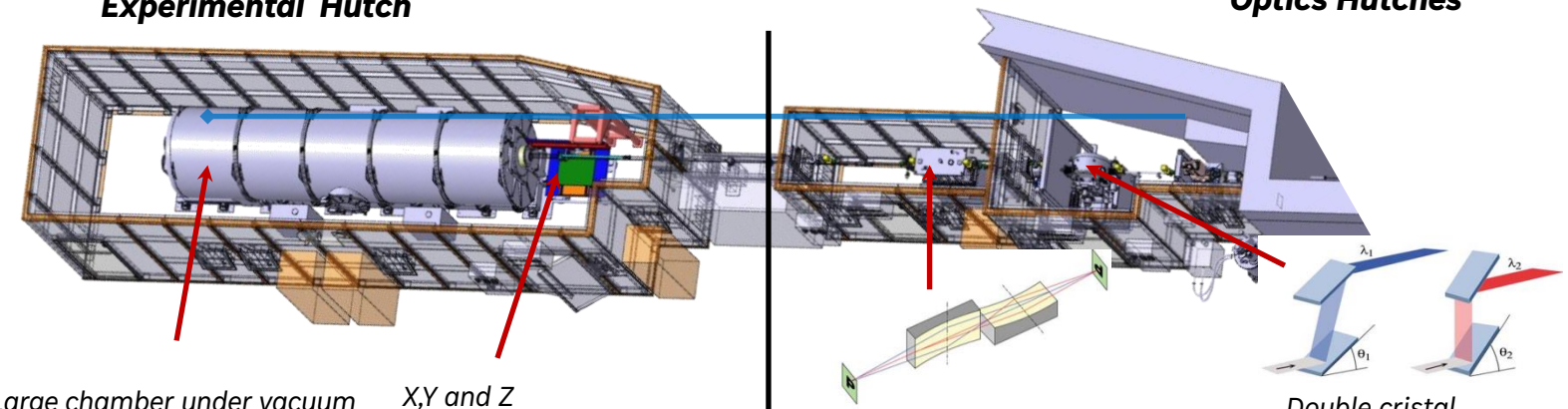
- A scattering pattern can be accurately calculated from atomic coordinates, thereby providing a link between high resolution and BioSAXS work.
- Using SAXS patterns, a*b initio* methods can propose « possible » shapes of a macromolecule
- Based on all-atom partial structural information, SAXS may constrain stochastic algorithms to provide « possible » tertiary configurations of multi-domain proteins
- Analysis and modeling require a monodisperse and ideal solution, which has to be checked independently.

The fact that your model fits the SAXS data does not prove that your model is correct.

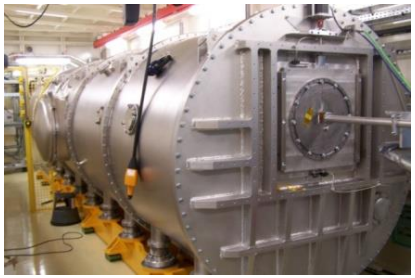
SAXS is at its best when used to discriminate preconceived hypotheses.

SWING Beamline

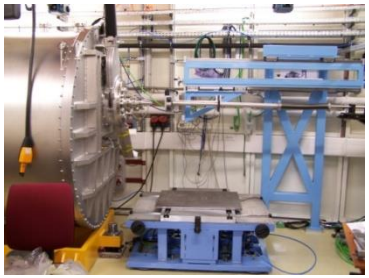
Experimental Hutch **Optics Hutches**



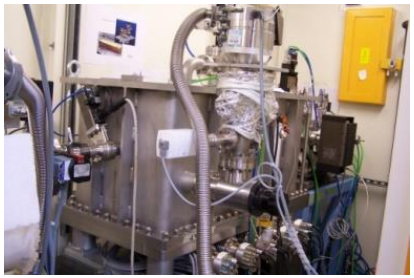
Large chamber under vacuum housing the X-rays detector



X,Y and Z motorized table



Two mirrors (KB)

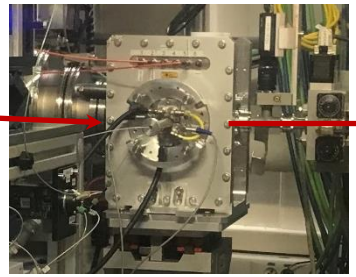


Double cristal Monochromator

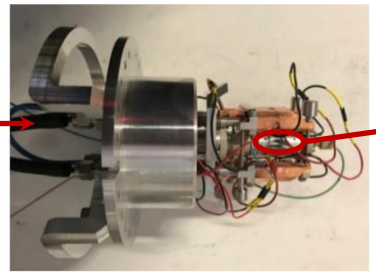
Sample environment dedicated to BioSAXS :

SEC-HPLC device

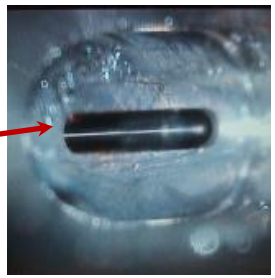
SEC-SAXS



SAXS cell vacuum chamber



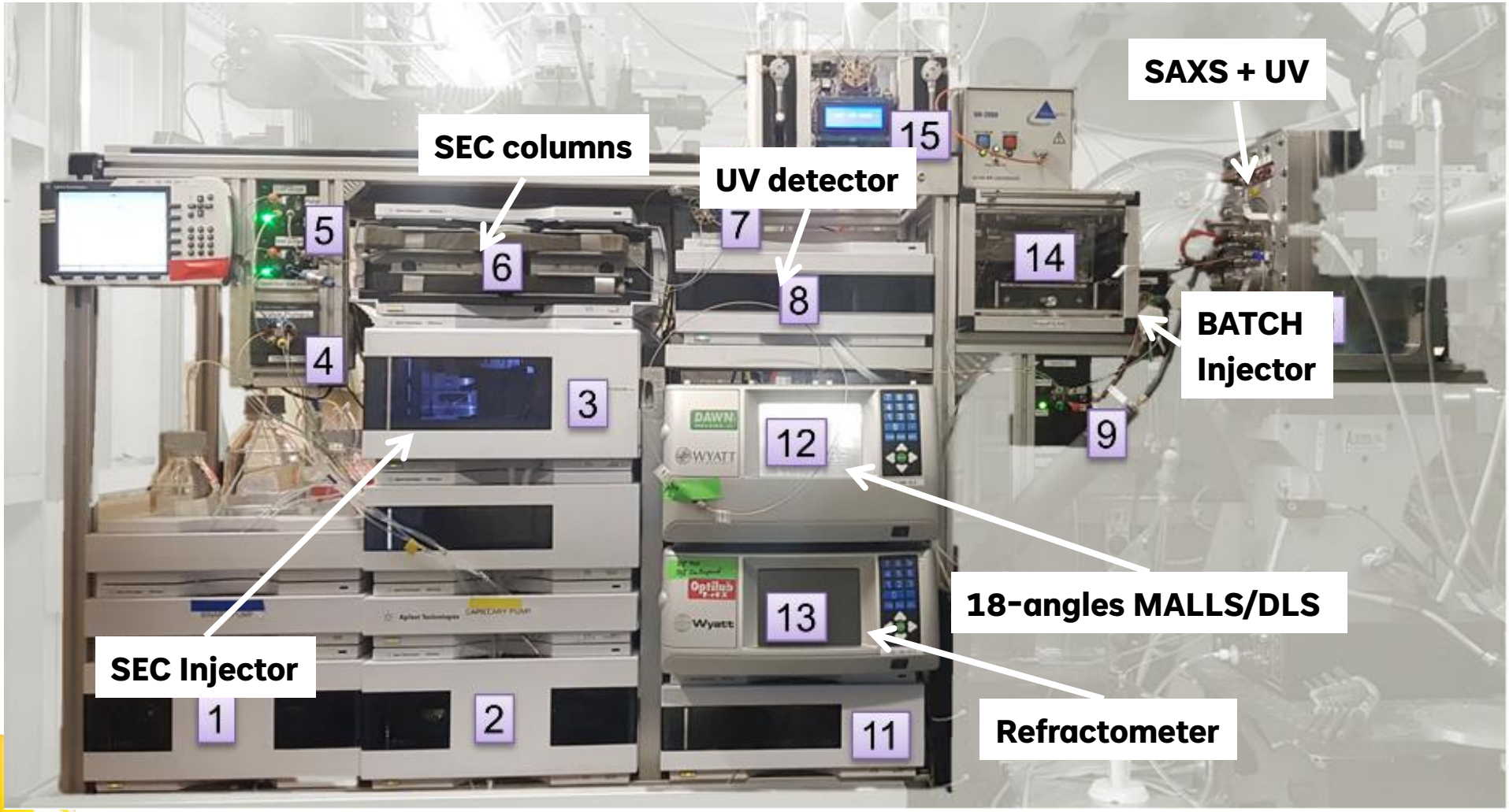
Details of the SAXS cell



Quartz capillary

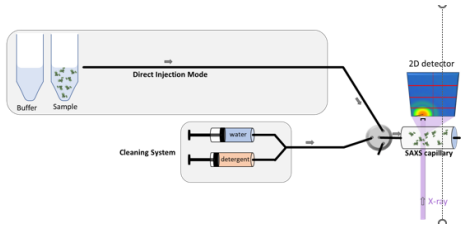
HPLC – MALLS/QELS – SAXS – RI online

Thureau, A., Roblin, P., & Pérez, J. (2021). *J. Appl. Cryst.*, *54*(6), 1698-1710.



SWING: a robotic arm for BioSAXS

- High-throughput injecting robot
  - Duty cycle : 3'30''
    - Injection / measurement / cleaning / drying
  - Injection volumes 10 to 50 µL
  - No dilution effect : [0.2 – 2 mg/mL]



~ 170 µL of tubing from injection to SAXS capillary

- Wetting effect induce loss of sample volume
- Difficulties for injecting viscous samples
- Cleaning and drying take time
- Cross contamination might occur

0 µL of tubing from injection to SAXS capillary

- No wetting effect
- Faster Cleaning and Drying process
- Viscous samples are allowed
- No cross contamination

ISAC SOLEIL Support Group (Laura Muñoz)

DeepSAXS

Dadimodo Implementation

- Initial (slow) version : Evrard et al. (2011), *J. Appl. Cryst.*, 44:1264-1271.
- Current (faster) version : O. Roudenko , A. Thureau, J. Pérez (2019), *GECCO '19, ACM*, NY, USA, 401-402.
  - Parallel implementation of the genetic algorithm
    - 7300 Atoms → 7 hours on a 20 processor node (200 generations)
  - User-friendly input
    - Tools for completion of pdb input files (if needed)
    - User-defined topology : Pdb file + rigid bodies definitions
  - Web server since end 2018
    - Accessible to external users (after login in Soleil DB)
    - Five independent runs launched in parallel

Before launching Dadimodo: PDB generation step

A. Sali & T.L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815, 1993.

Shell script launching « Modeller »

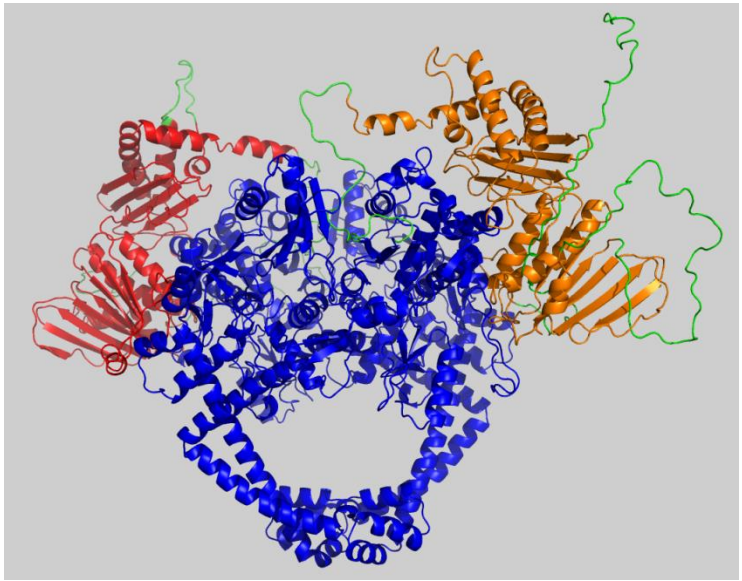
A script is available on Swing Website
https://www.synchrotron-soleil.fr/en/beamlines/swing

Original PDB files
missing atoms & residues

FASTA sequence of the entire protein

Generates

- missing atoms coordinates
- missing residues in linkers and tails
- aleatory orientations for flexibly connected domains



Mycobacterium tuberculosis DNA Gyrase (PDB 6GAV)

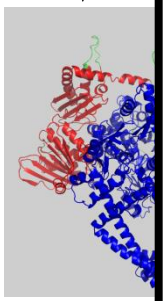
Complete PDB file
Interpretable by MMTK, the MD python library used in Dadimodo

Dadimodo Web Server

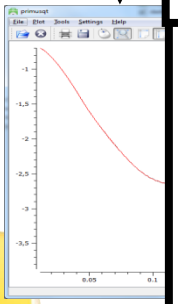
https://dadimodo.synchrotron-soleil.fr

3 input files needed to launch Dadimodo on the Web Server

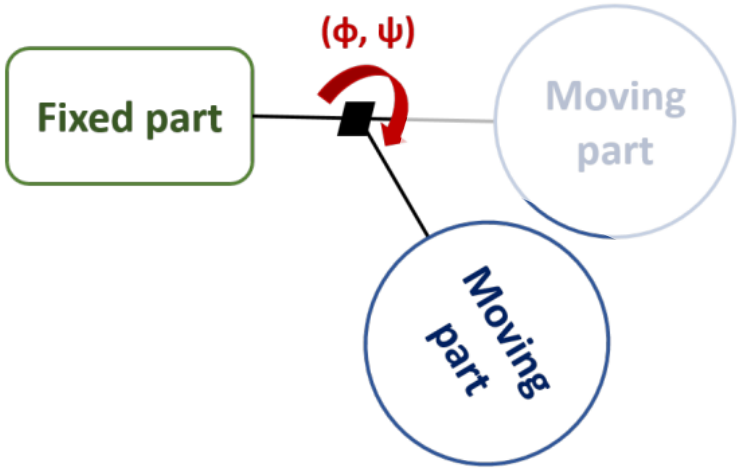
Complet



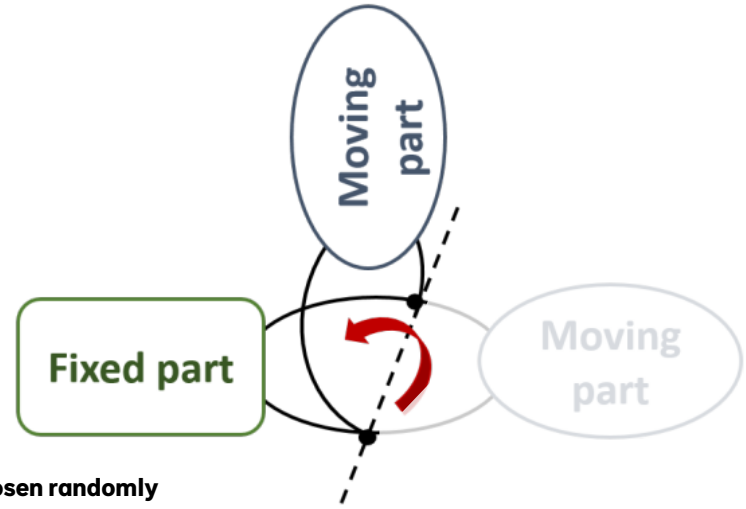
SAX



Linker (ϕ, ψ) mutation



Double linker (axis) mutation



Mutated residue is chosen randomly

B: 83 – 124

loop

B: 427 – 454

linker

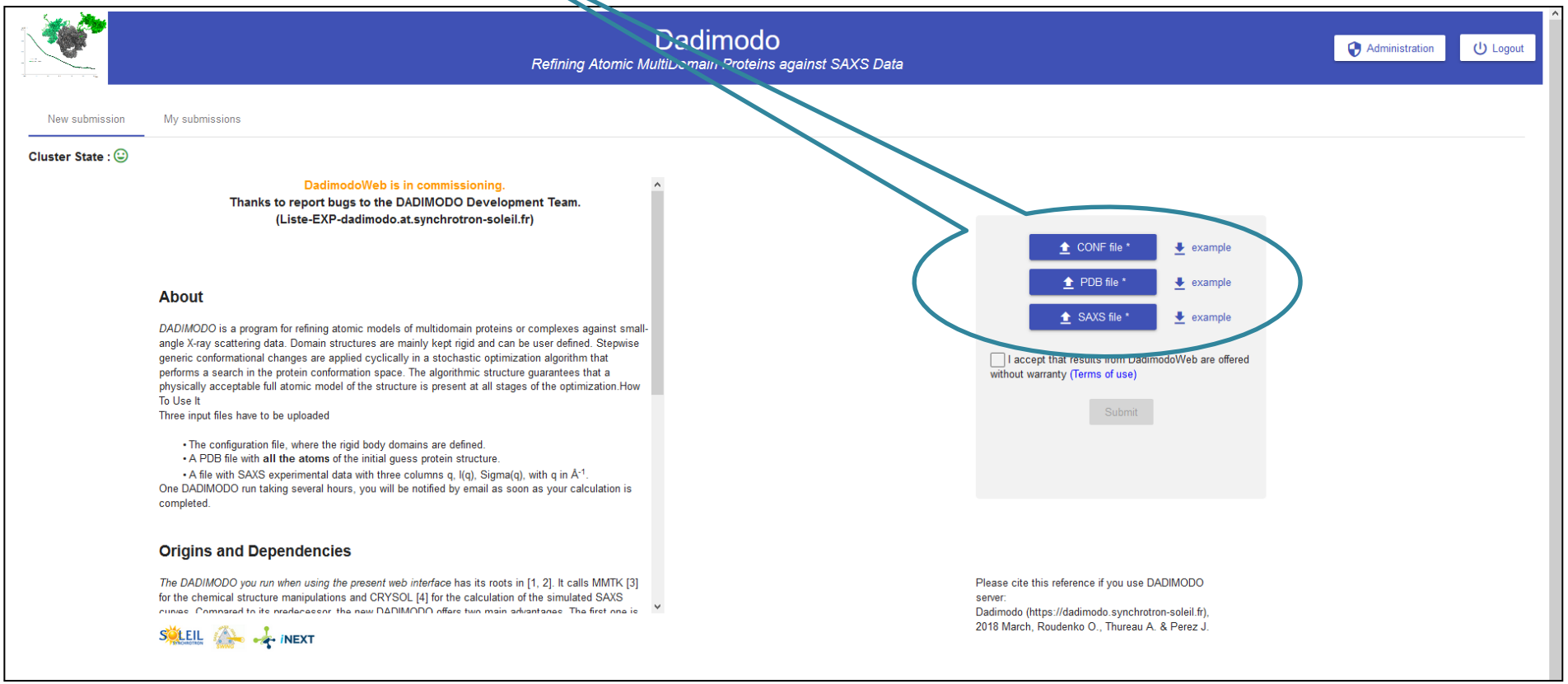
A: 422 - 452

A: 453 – 1179
&
B: 455 – 1179

Dadimodo Web Server

https://dadimodo.synchrotron-soleil.fr

3 input files needed to launch Dadimodo on the Web Server

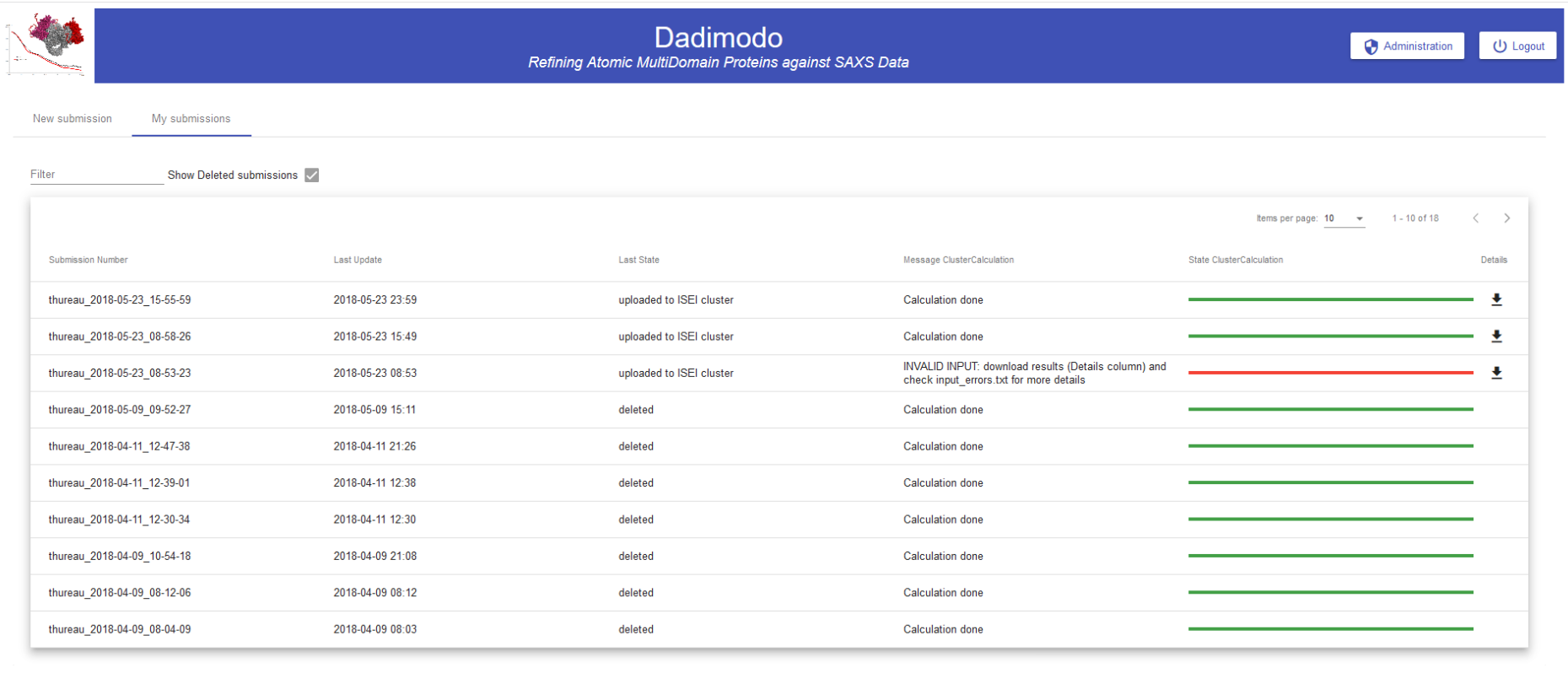


Dadimodo Web Server

https://dadimodo.synchrotron-soleil.fr

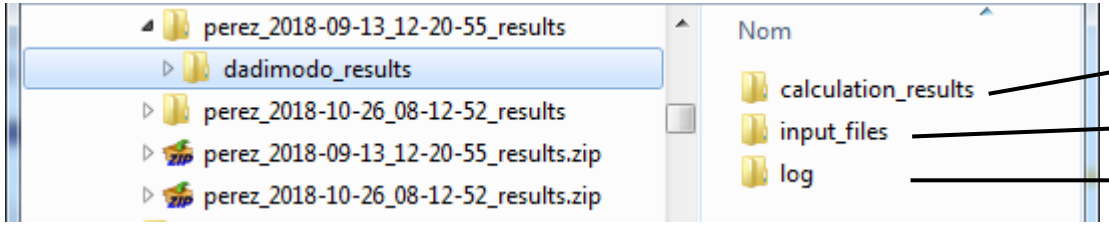
« My submissions » tab:

- ***Status of current submission and history of past jobs***
- ***Results download (zip file)***



Dadimodo results files

https://dadimodo.synchrotron-soleil.fr



The final pdb files from the several runs

The 3 input files

The exhaustive log files + the statistics files