Solution X-ray Scattering from Biological Macromolecules

Javier Pérez

Beamline SWING, Synchrotron SOLEIL, Saint-Aubin, France

General Outline

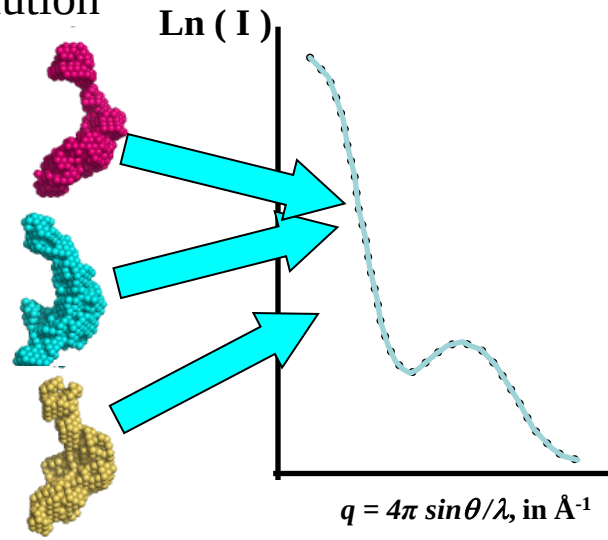
- Introduction
- SAXS basics
- Data Analysis
- A few experimental considerations
- Modelling
- Conclusions

INTRODUCTION

Principles of Small Angle X-ray Scattering in solution

SAXS provides structural information about macromolecules in solution

- **Limits**
  - spherically averaged information → 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

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

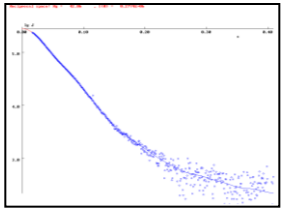
3D structural modeling

- 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

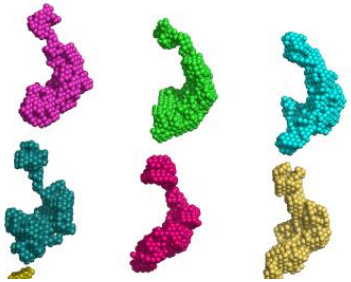
→ Models « compatible with SAXS data »
NOT unique models, NO electronic density maps.

Structural information about macromolecules in solution

Nothing known (except the curve)

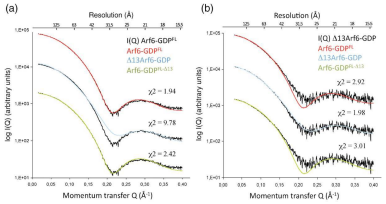


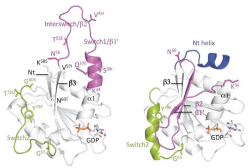
Shape determination



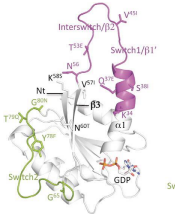
DAMMIN
DAMMIF
DENFERT

Known or supposed all-atom models



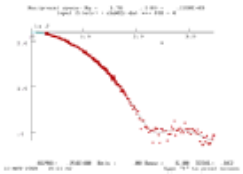


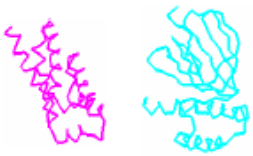
Model validation / elimination



CRYSOL
FOXS

Structure of subunits available

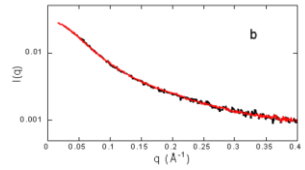


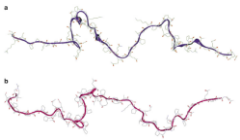


Rigid body modeling of the complex

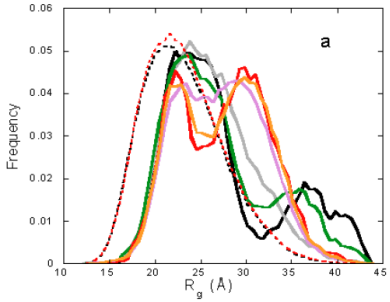
SASREF
BUNCH
CORAL
DADIMODO

Zones of supposed high flexibility





Selection within an Ensemble of Random Conformations

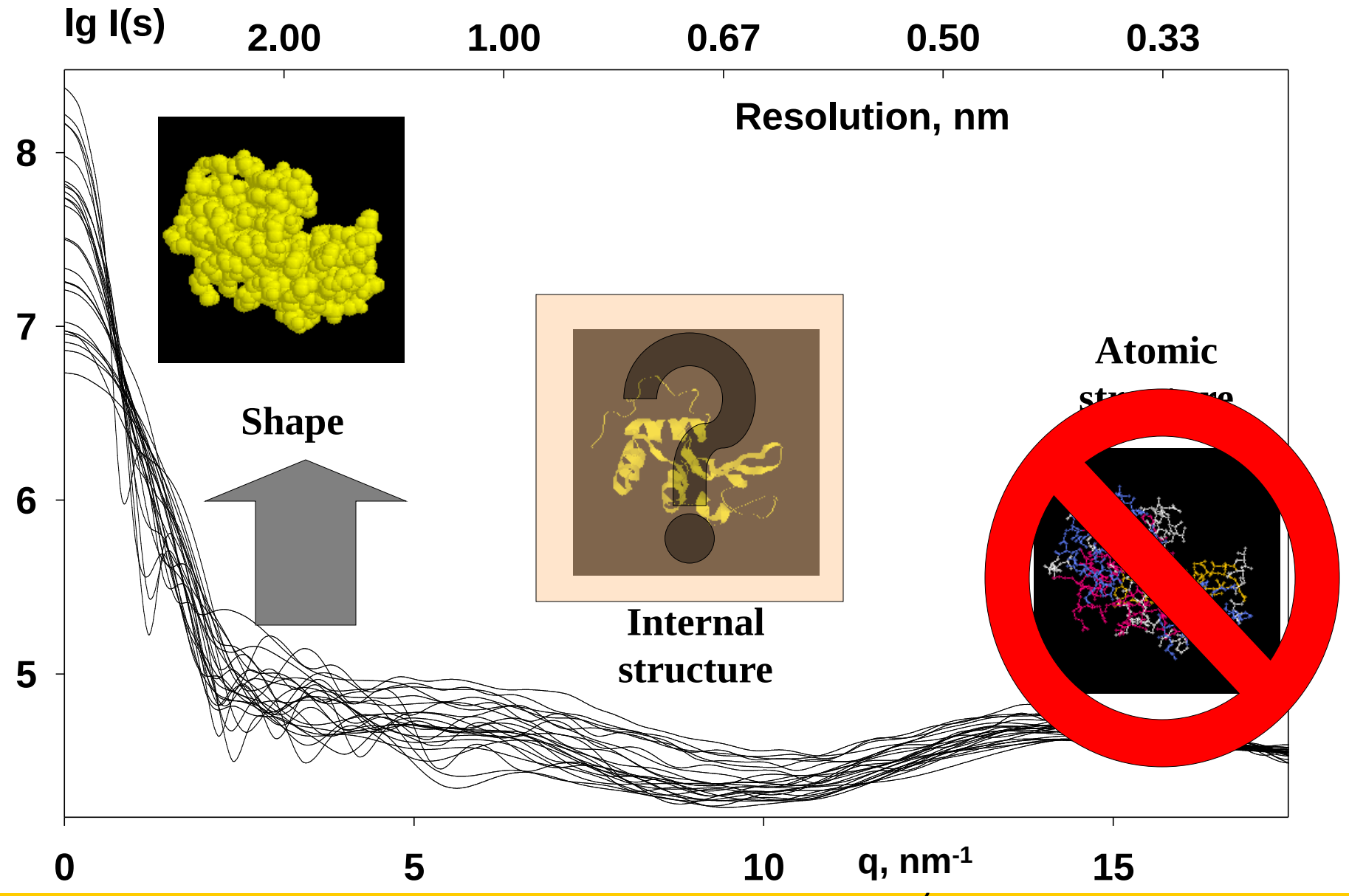


EOM

MES

Slide from Dmitri Svergun, EMBL Hamburg

What may solution scattering yield?



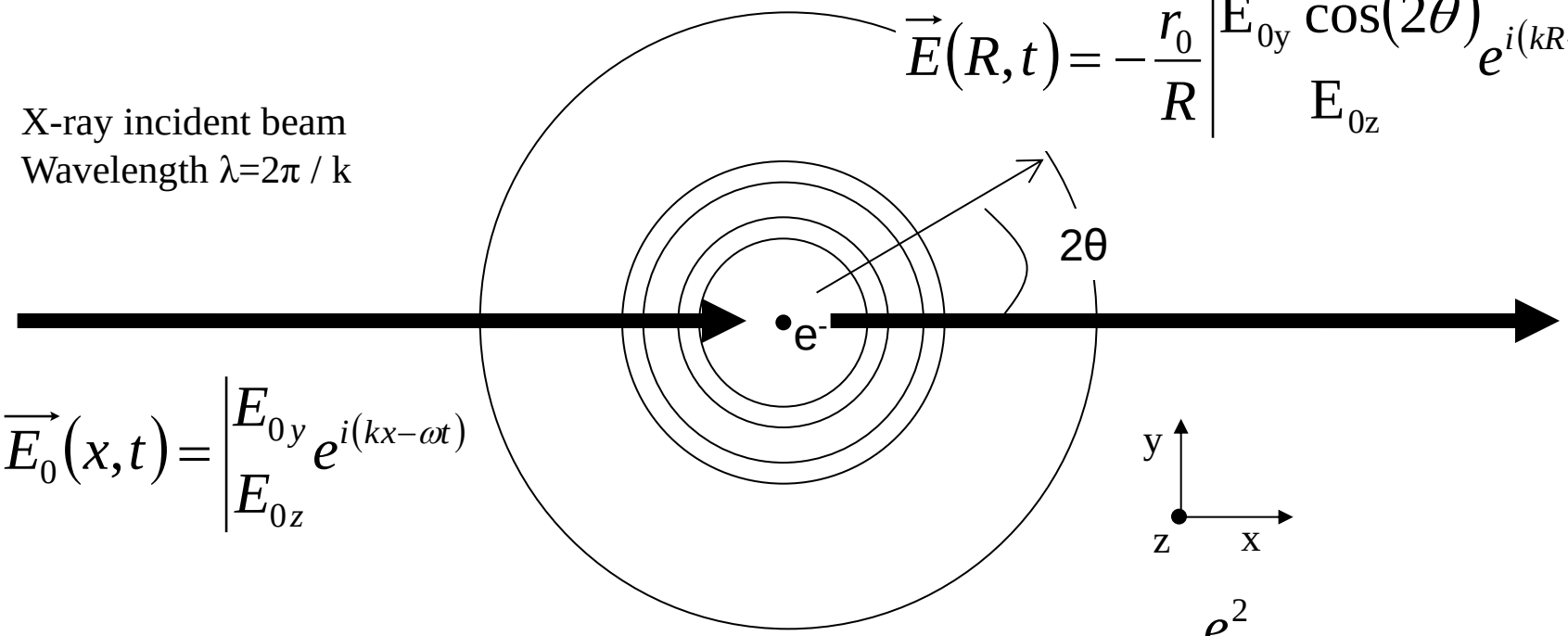
SAXS BASICS

Elastic Thompson scattering by an electron

- **What scatter X-rays are <u>the electrons</u>**

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

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



$$\vec{E_0}(x,t) = \begin{vmatrix} E_{0y} \\ E_{0z} \end{vmatrix} e^{i(kx-\omega t)}$$

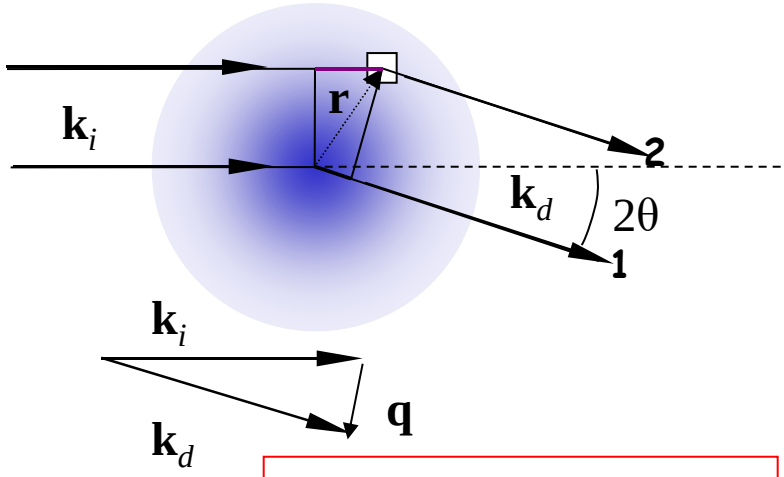
$$r_0 = \frac{e^2}{4\pi\varepsilon_0 mc^2} = 2{,}818\ 10^{-15}\, m$$

- **r_0 is called the electron classical radius**

Scattering amplitude by a a particle

Coherent scattering : summing up amplitudes

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



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

$$Wave\,2: \overrightarrow{E_2}(R,t) = -\frac{r_0}{R}\overrightarrow{E_0}(2\theta)e^{i\left(kR-\omega t+\vec{k}_i\cdot\vec{r}-\vec{k}_d\cdot\vec{r}\right)}$$

Phase shift between waves 1 and 2 :

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

$\vec{q}$ = Momentum transfer

$$q = \left\|\vec{q}\right\| = \frac{4\pi\sin(\theta)}{\lambda}$$

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

Particle scattering « length » (or Amplitude):

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

Intensity scattered by a sample – Auto-correlation function

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

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

$$I(\vec{q}) = \frac{1}{V} A.A^*(\vec{q}) = \frac{r_0^{\,2}}{V} \int_V \int_V \rho_e(\vec{r}_1) e^{-i\vec{q}\cdot\vec{r}_1} \rho_e(\vec{r}_2) e^{+i\vec{q}\cdot\vec{r}_2} d^3\mathbf{r}_1 d^3\mathbf{r}_2$$

$$I(\vec{q}) = \frac{r_0^{\,2}}{V} \int_V \int_V \rho_e(\vec{r}_1)\rho_e(\vec{r}_2) e^{-i\vec{q}\cdot\left(\vec{r}_1-\vec{r}_2\right)} d^3\mathbf{r}_1 d^3\mathbf{r}_2$$

Auto-correlation function $\gamma_e(\mathbf{r})$:

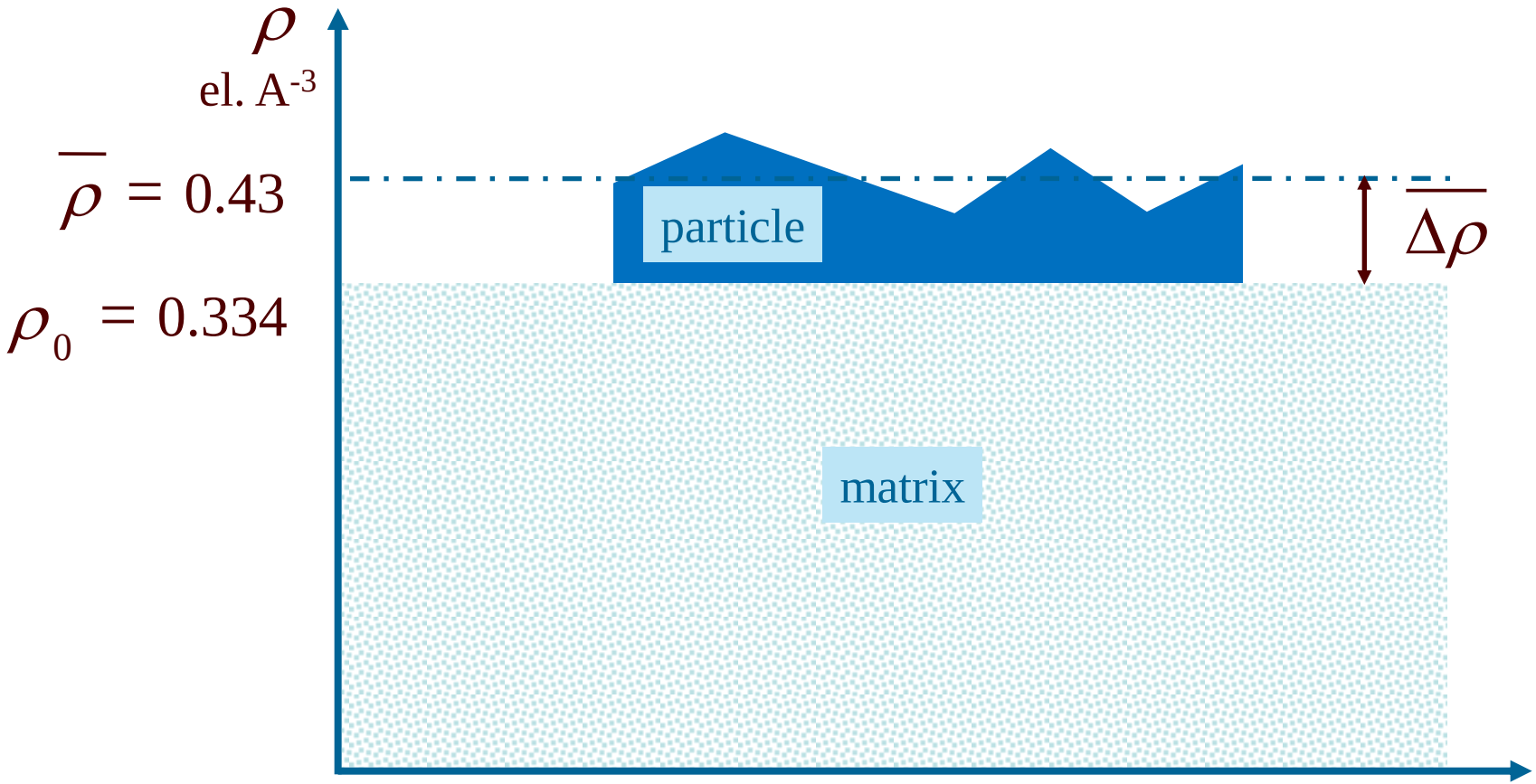
$$\gamma_e\left(\vec{r}\right) = \frac{1}{V} \int_V \rho_e(\vec{r}')\rho_e(\vec{r}+\vec{r}') d^3\mathbf{r}'$$

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

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



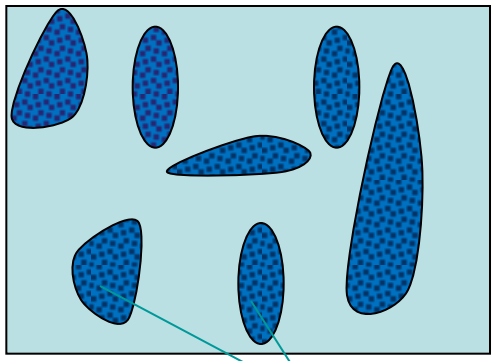
Particles inserted in a "matrix"

- Scattering amplitude

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

Particles volume

Electronic Density Contrast



Particles

- $\Delta\rho(\vec{r})$ 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

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

Irradiated volume

- I(**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:

$$I_1(q) = \overline{\left\langle f_1\left(\vec{q}\right) f_1^{*}\left(\vec{q}\right)\right\rangle_{\Omega}}$$

Modulus

Vector

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

Average Electronic

Particle volume

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

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

Basic law of reciprocity in scattering

All, including **large**, distances Δr in the particle ⟺ **Small** scattering angle q

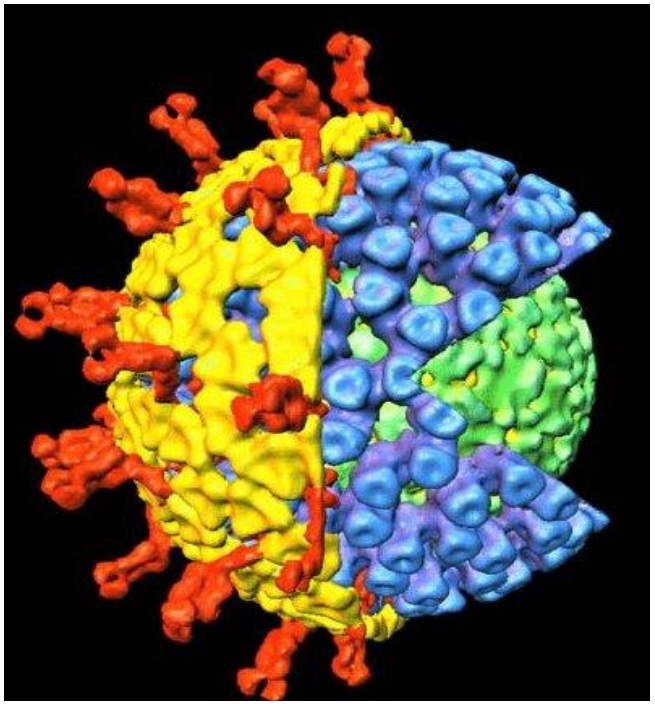
Short distances Δr in the particle ⟺ **Large** scattering angle q

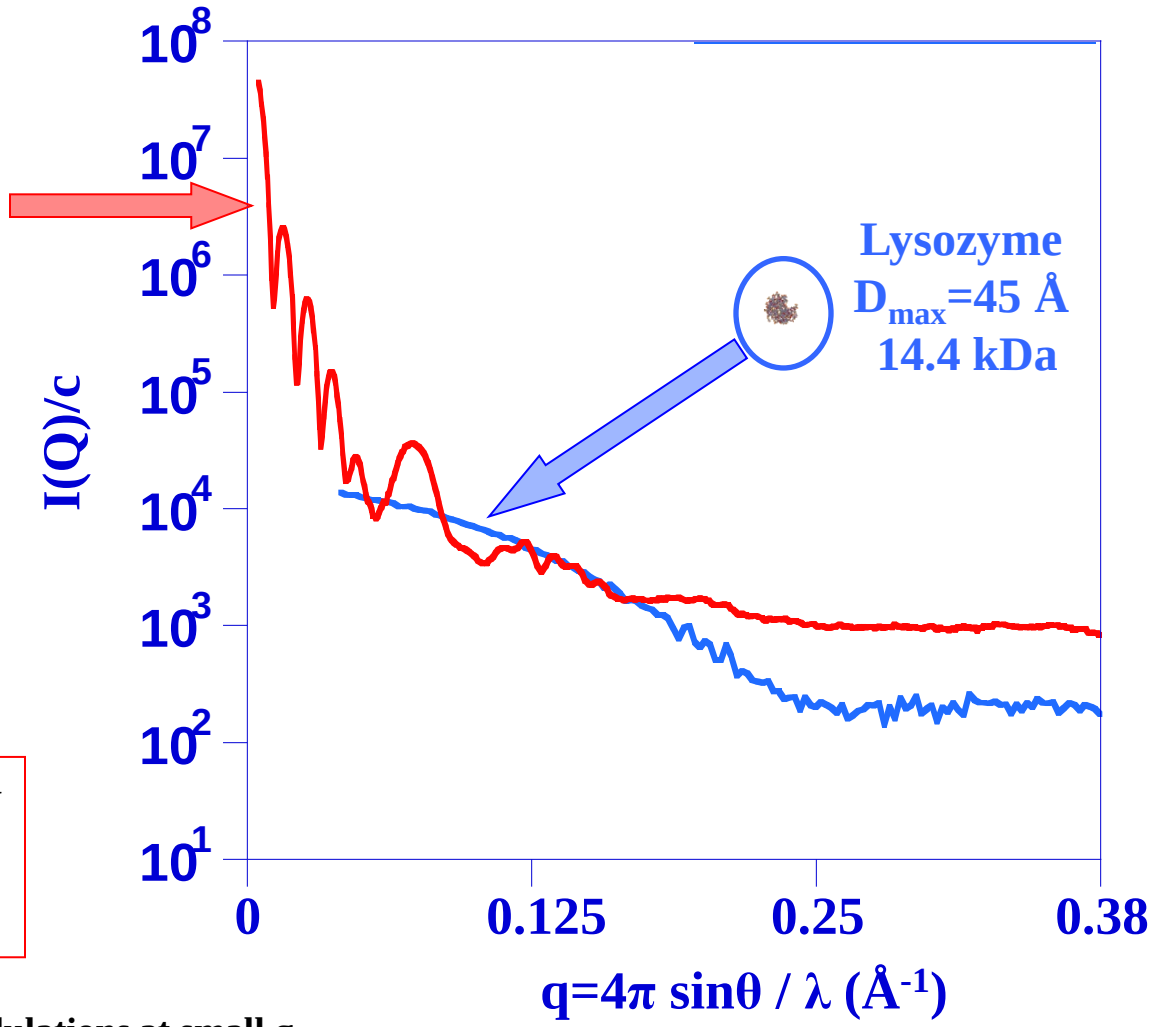
Phase : **q·Δr**

$$I(\vec{q}) = \frac{r_0^{\,2}}{V} \int_{V_1}\int_{V_2} \Delta\rho_e(\vec{r}_1)\Delta\rho_e(\vec{r}_2) e^{-i\vec{q}\cdot\left(\vec{r}_1-\vec{r}_2\right)} d^3\mathbf{r}_1 d^3\mathbf{r}_2$$

Basic law of reciprocity in scattering

Rotavirus VLP : diameter = 750 Å, 44 MDa

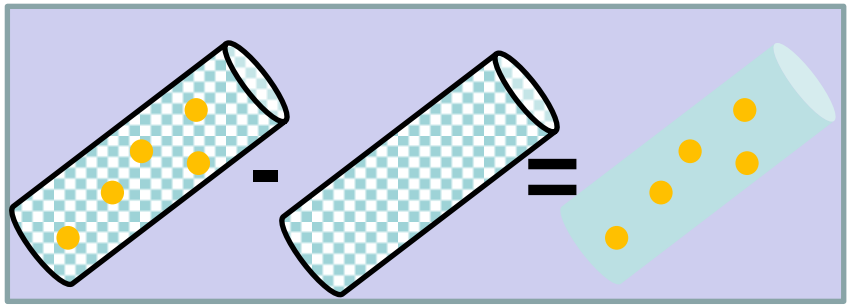




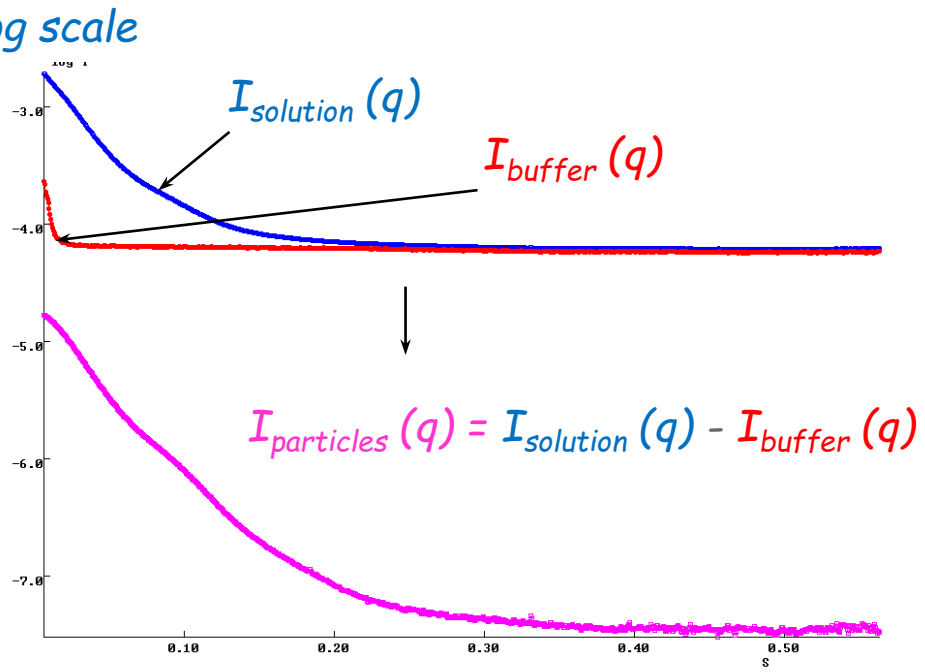
Typical range on beamline SWING
$2\ 10^{-3} < q < 7\ 10^{-1}$ Å^{-1}
750 Å > D_{max}
3100 Å > d_{Bragg} > 9 Å

Long distance correlations ⟺ modulations at small q

A SAXS curve results from a pair of measurements : solution & buffer



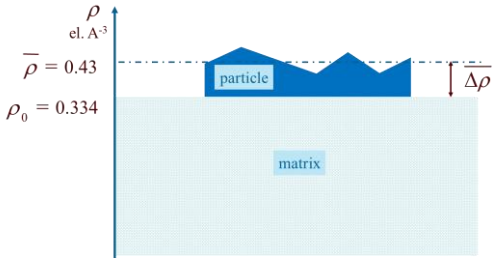
$I_{solution}(q)$ - $I_{buffer}(q)$ = $I_{particles}(q)$



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.

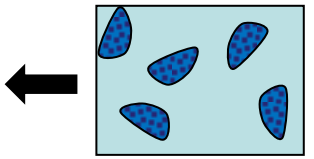
Do not get mixed up !

contrast effect

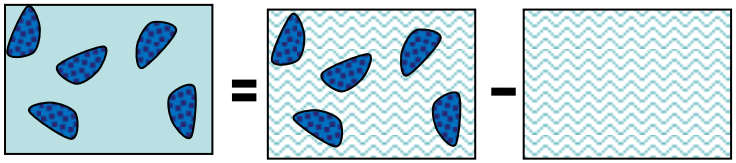


$$\Delta\rho(\vec{r}) = \rho(\vec{r}) - \rho_0$$

$$f(\vec{q}) = -r_0 \int_V \Delta\rho(\vec{r}) e^{-i\vec{q}\cdot\vec{r}} d^3\mathbf{r} \quad \rightarrow \quad I(q) = \overline{\langle f(\vec{q}) f^*(\vec{q}) \rangle_\Omega}$$



buffer subtraction



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

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

- **Ideal**

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

- **Monodisperse** $i_i(q) = i_1(q)$ whatever i

- **Ideal and monodisperse**

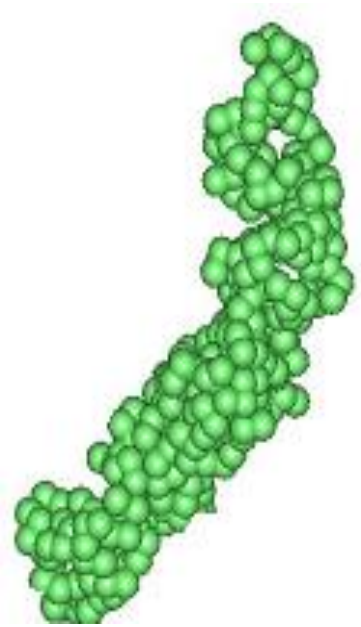
$$I(q) = \mathrm{N}\, i_1(q) = N \left\langle f_1\left(\vec{q}\right) f_1^{*}\left(\vec{q}\right) \right\rangle_{\Omega}$$

$i_1(q)$

Ideality

Monodispersity

$I(q)$



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

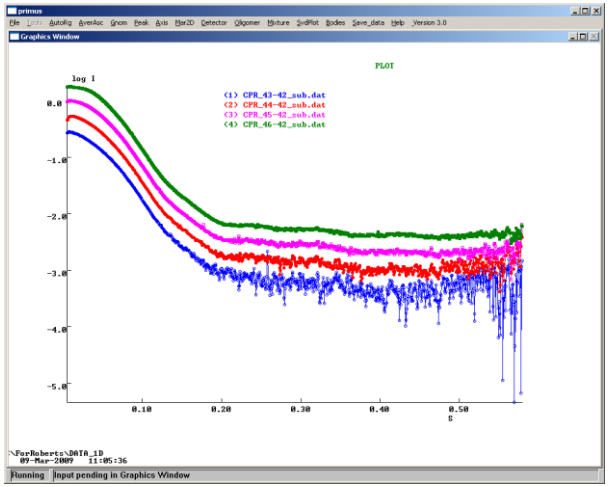
molecule

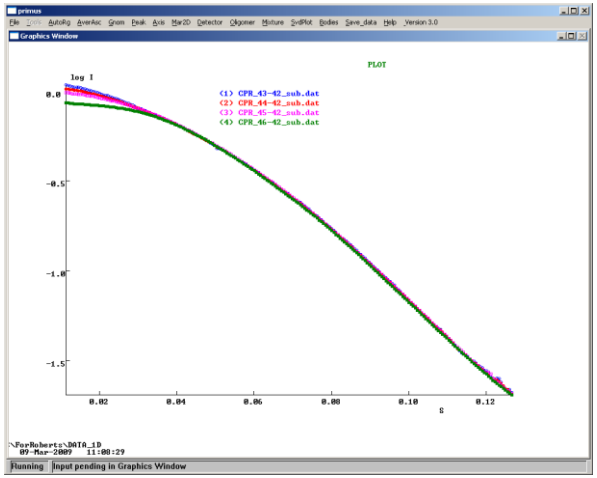
experimental

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.**

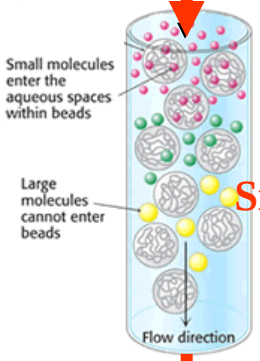




SE-HPLC / Solution Sampler

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

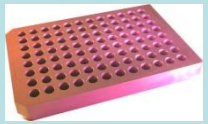


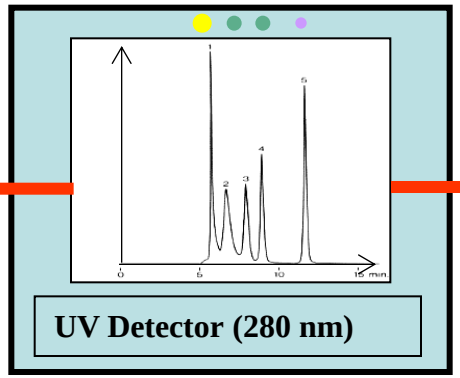
Size Exclusion

Incident X-ray

Pump

Injection-mixing





UV-Vis in

SAXS Cell

UV-Vis out

Flow rate 5-40 µl/min

Pure sample

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

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 / Rg for a globular shape.

Prof. André Guinier
1911-2000
Orsay, France

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

Extrapolated intensity at origin

Radius of gyration

Guinier law, in Log scale :

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

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).

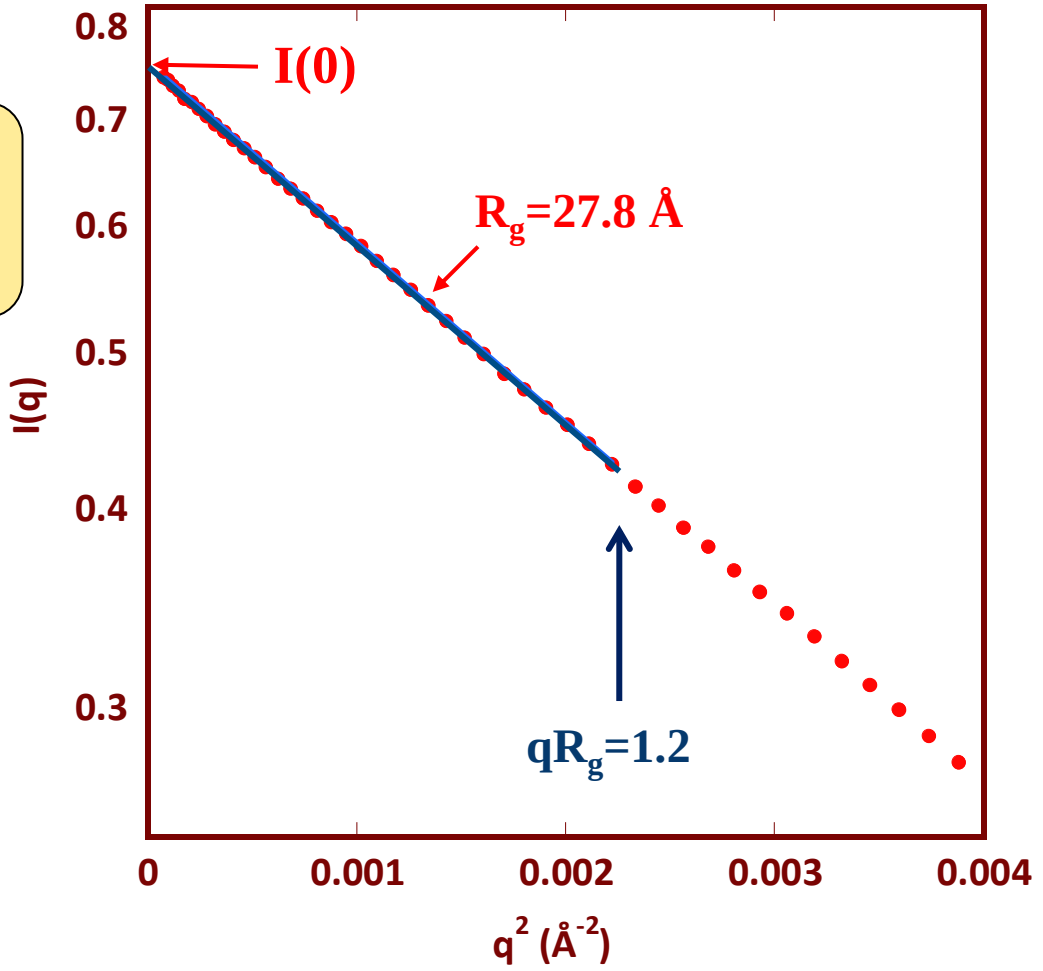
Data Analysis : Guinier law

Guinier analysis

R_g → size

I(0) → mol mass / oligomerisation state

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



Validity range :

$0 < R_g q < 1$ for a solid sphere

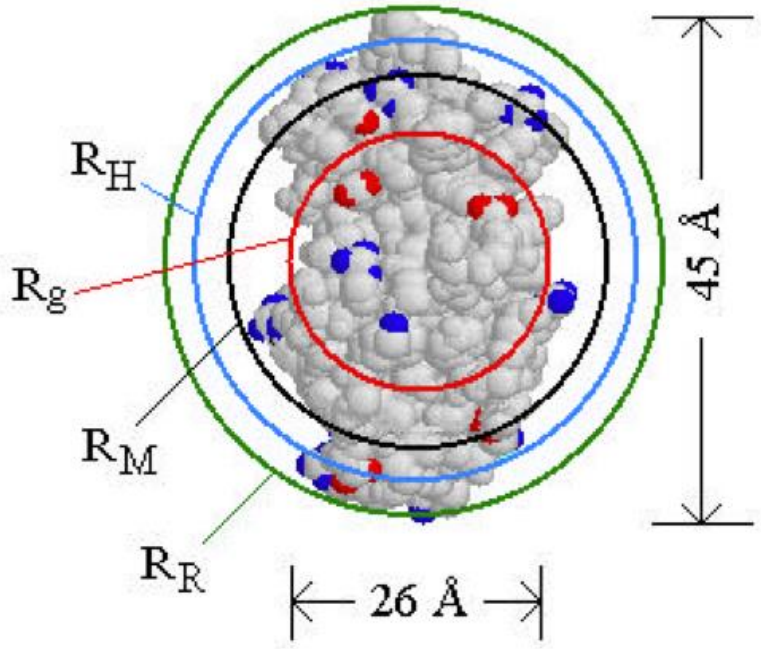
$0 < R_g q < 1.3$ rule of thumb for a globular protein

ideal monodispersed

Radius of gyration

$$Rg_{\exp}^{2} = \frac{\int_V r^2 \Delta\rho(\vec{r})\, d\vec{r}}{\int_V \Delta\rho(\vec{r})\, d\vec{r}}$$

Lysozyme



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

Useful definitions of R_g

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

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

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

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

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

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

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

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

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, c, is precisely known …)**

Typically :

M (kDa) = 1500 * I0 (cm^{-1}) / C (mg/ml)

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

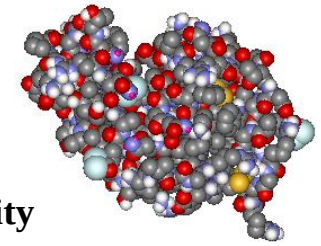
Bernado et al. (2009), Biophys. J., 97 (10), 2839-2845.

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**

M=14.3 kDa



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

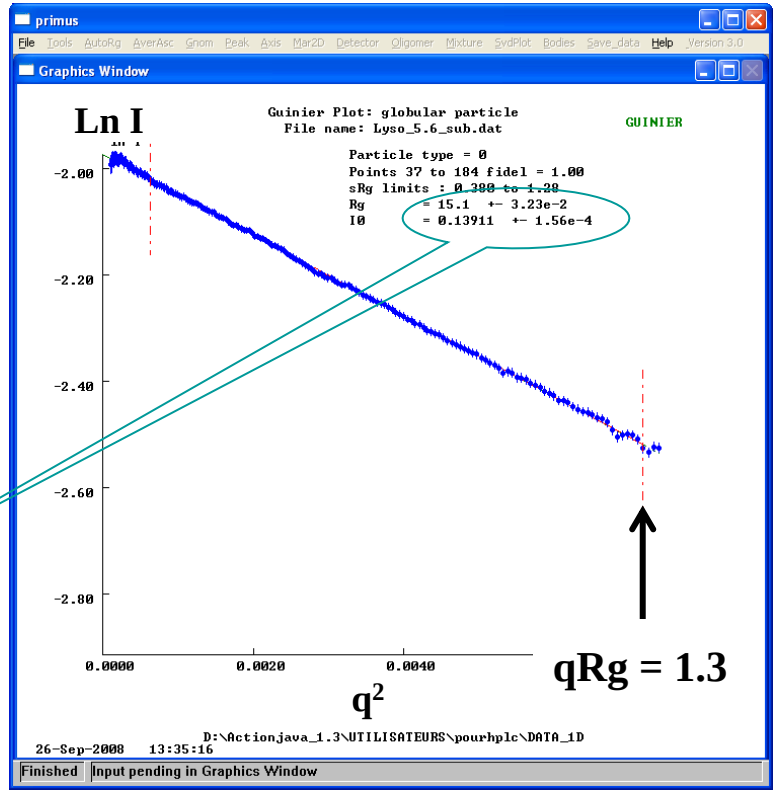
R_g = 15.1 ± 0.03 Å

$I_{exp}(0)$ = 0.0543 cm^{-1}

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

$M_{exp}(kDa) = I_{exp}(0) * 1500 / c,$

→ M_{exp} = 14.6 kDa



From Rg, supposing the protein is globular:

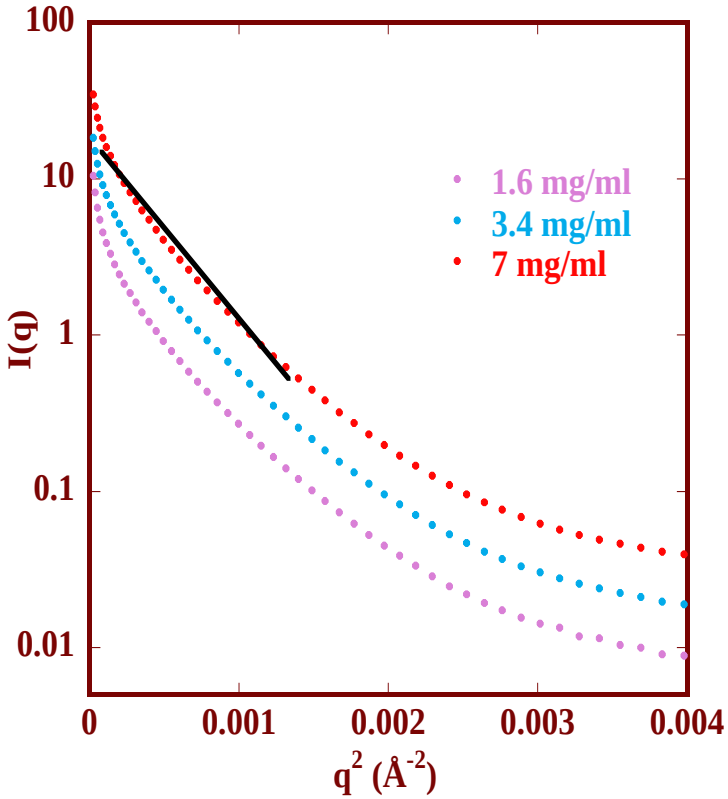
$M_{Rg}(kDa) = (Rg / 6.3)^3$

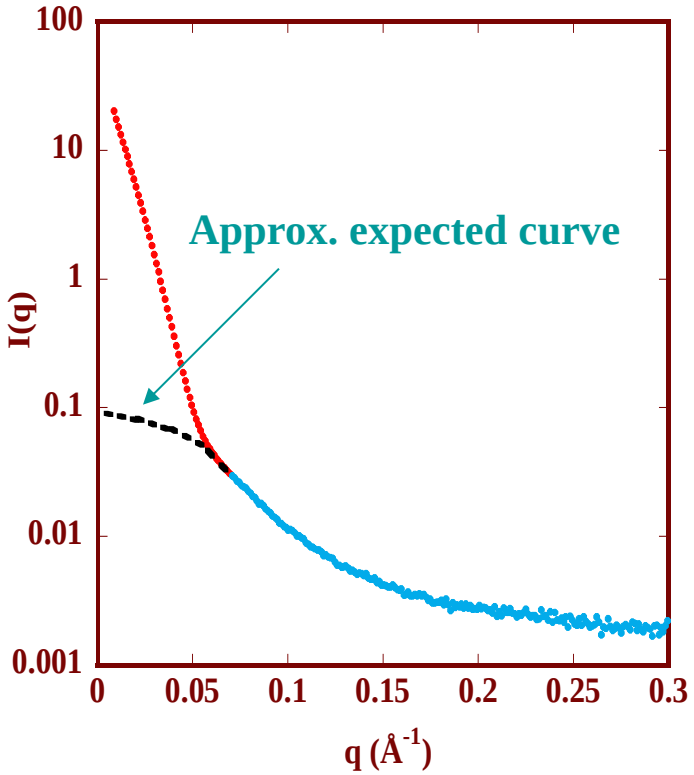
→ M_{Rg} = 13.8 kDa

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

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

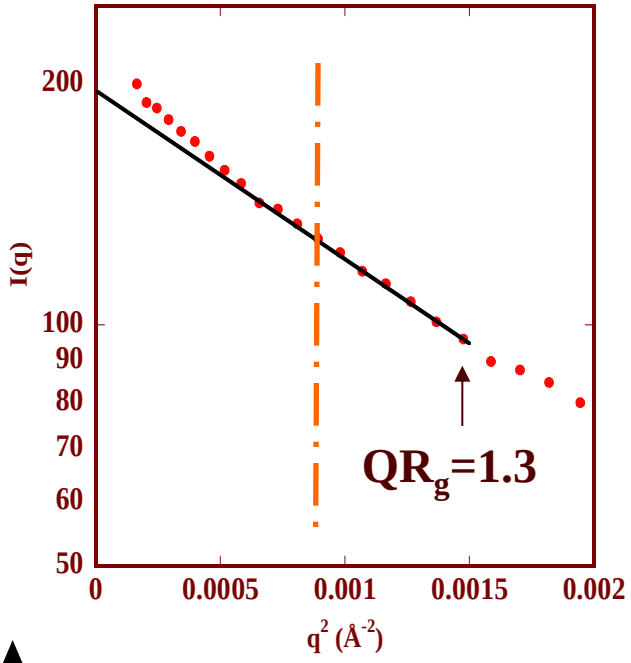
(Courtesy D. Durand, IBBMC, Orsay)

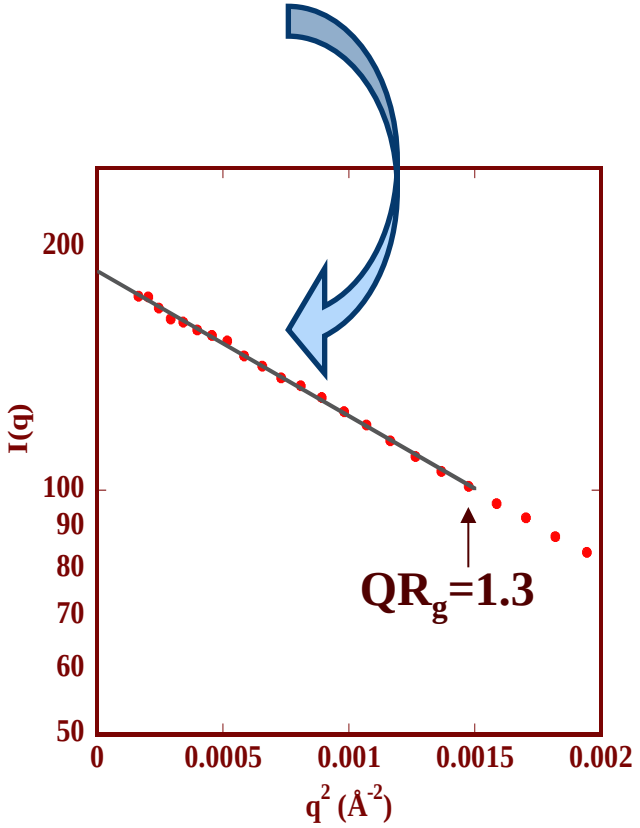
Evaluation of the solution properties

Weak aggregation → **possible improvement**

centrifugation, buffer change

Nanostar –PR65 protein





R_g ~ 38 Å – too high!!

R_g ~ 36 Å

(Courtesy D. Durand, IBBMC, Orsay)

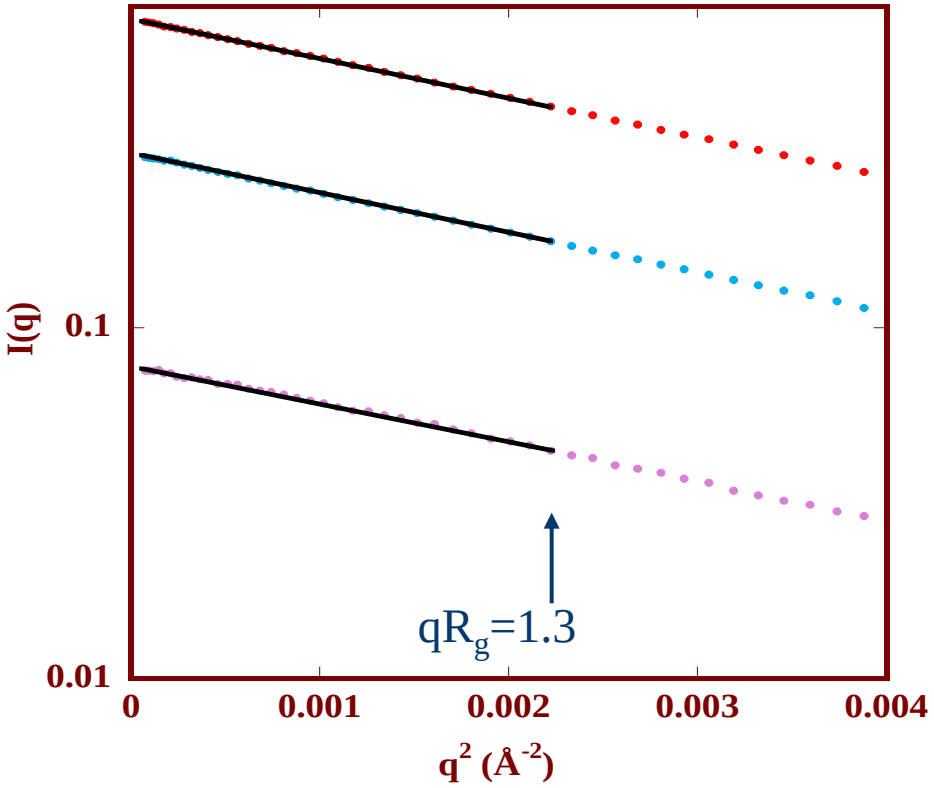
Evaluation of the solution properties

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



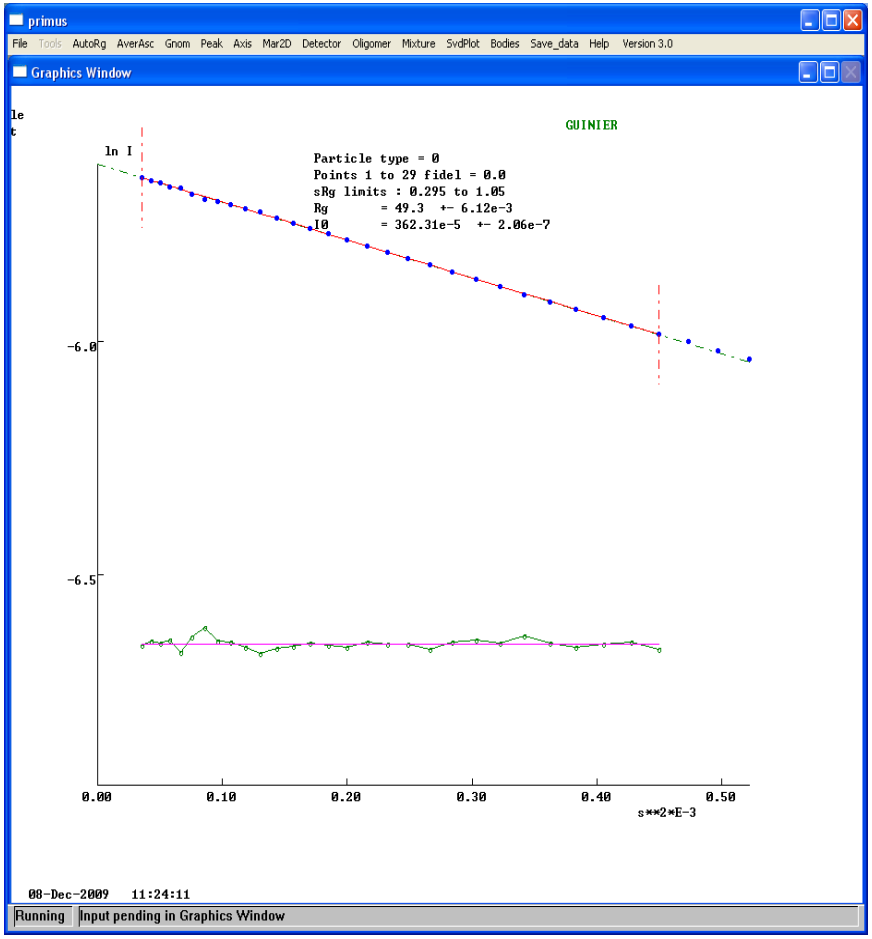
same R_g at all three concentrations

No interactions.

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

Evaluation of the solution properties

Guinier plot



c = 4

R_g = 49.3 Å

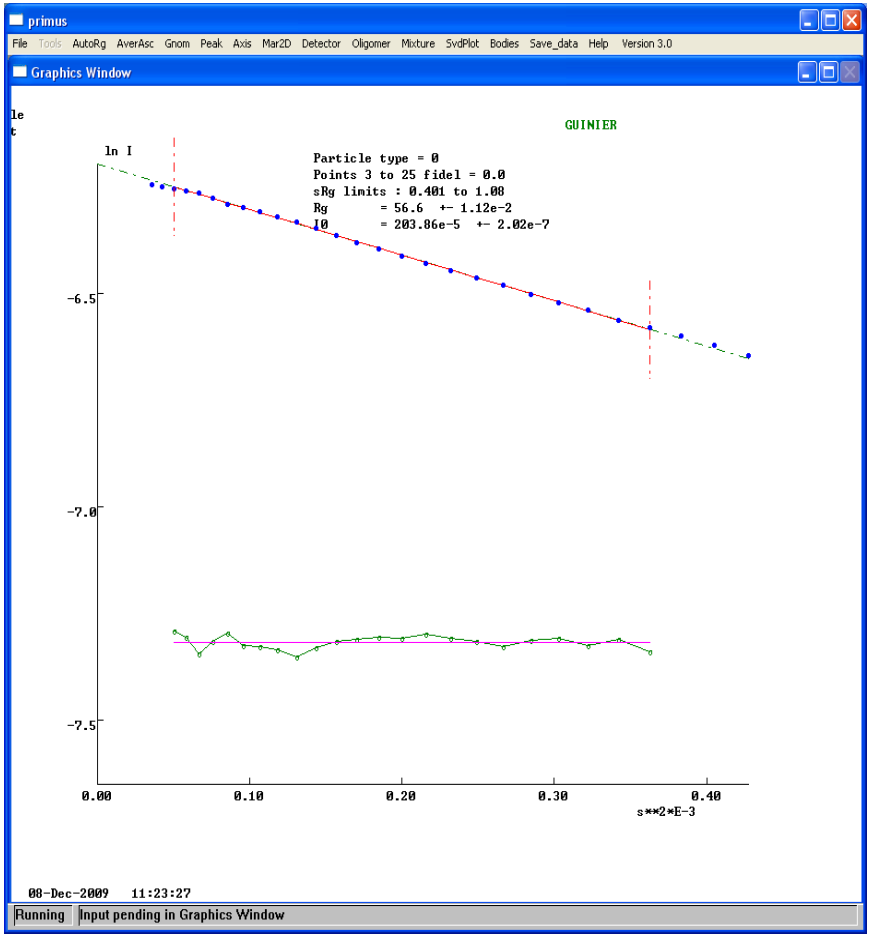
RNA molecule
L. Ponchon, C. Mérigoux *et al.*

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.*

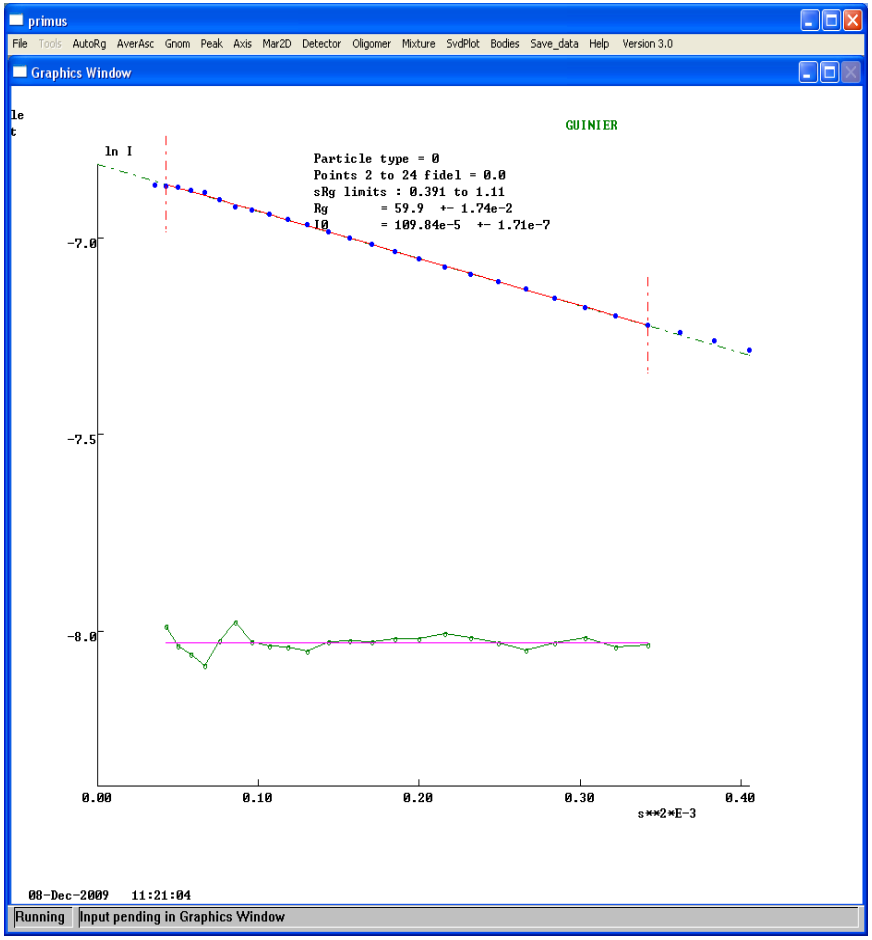
Evaluation of the solution properties

Guinier plot

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.*

Evaluation of the solution properties

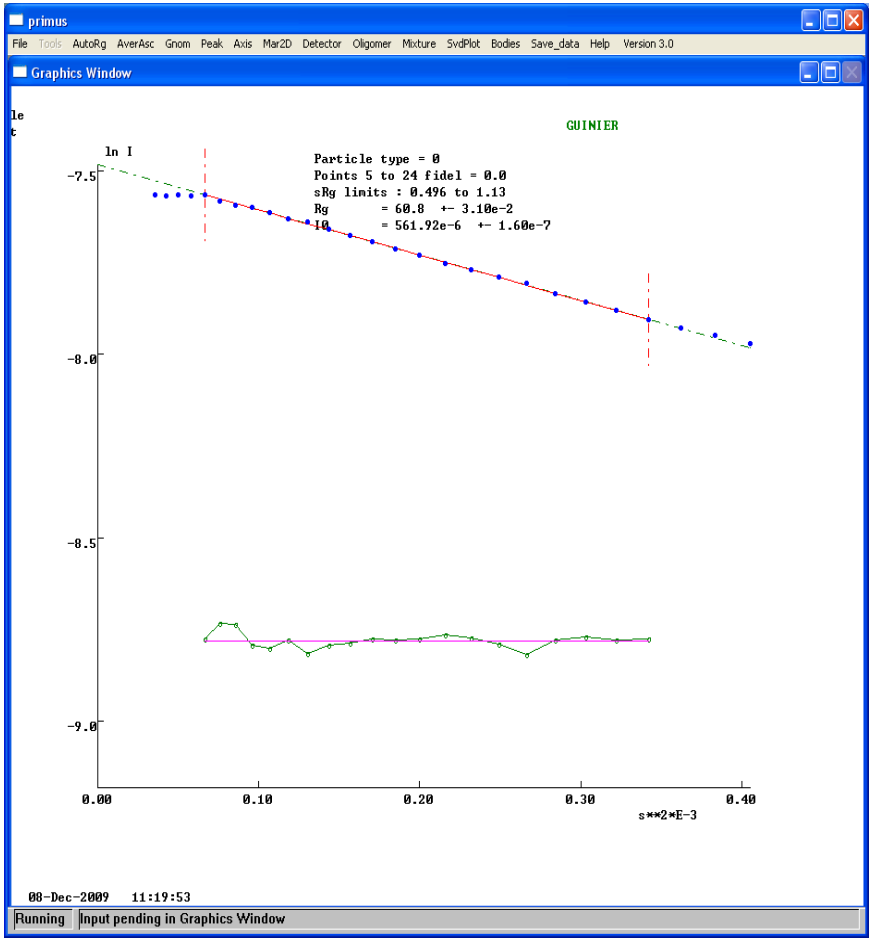
Guinier plot

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

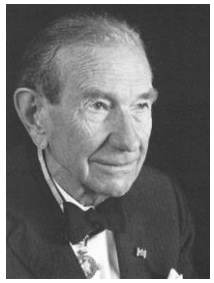


RNA molecule
L. Ponchon, C. Mérigoux *et al.*

Data Analysis

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

Kratky Plot

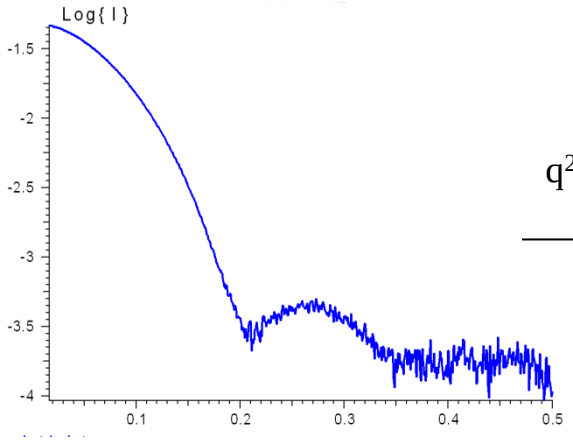


Prof. Otto Kratky
1902-1995
Graz, Austria

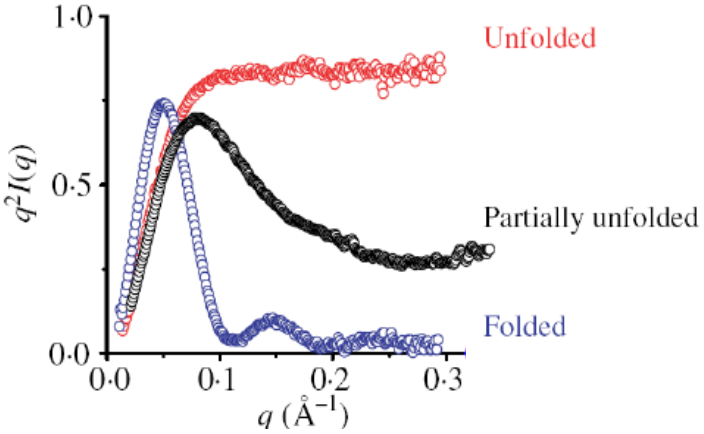
SAXS provides a sensitive means to ***evaluate the degree of compactness*** of a protein:

- 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



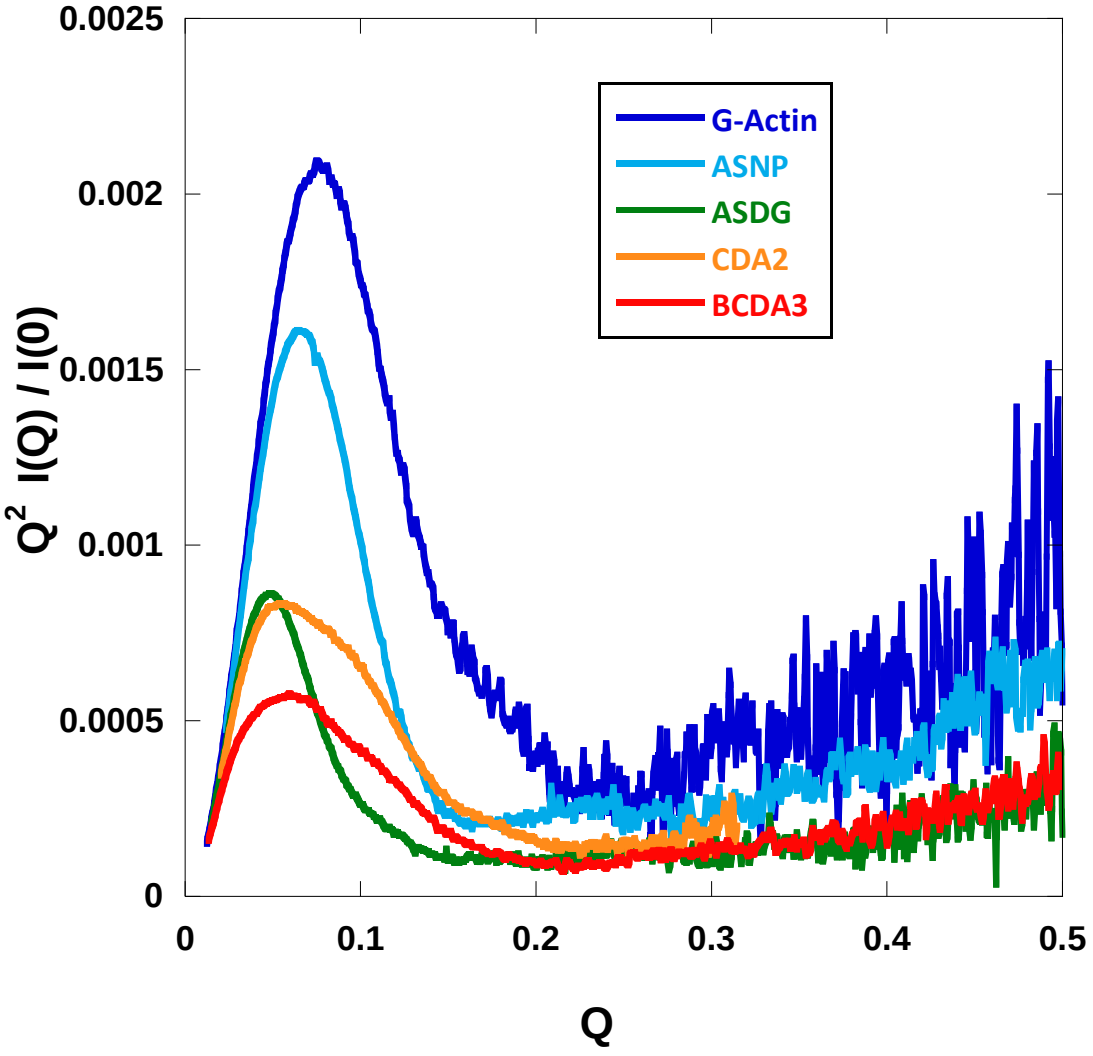
Putnam, D., et al. (2007) Quart. Rev. Biophys. 40, 191-285.

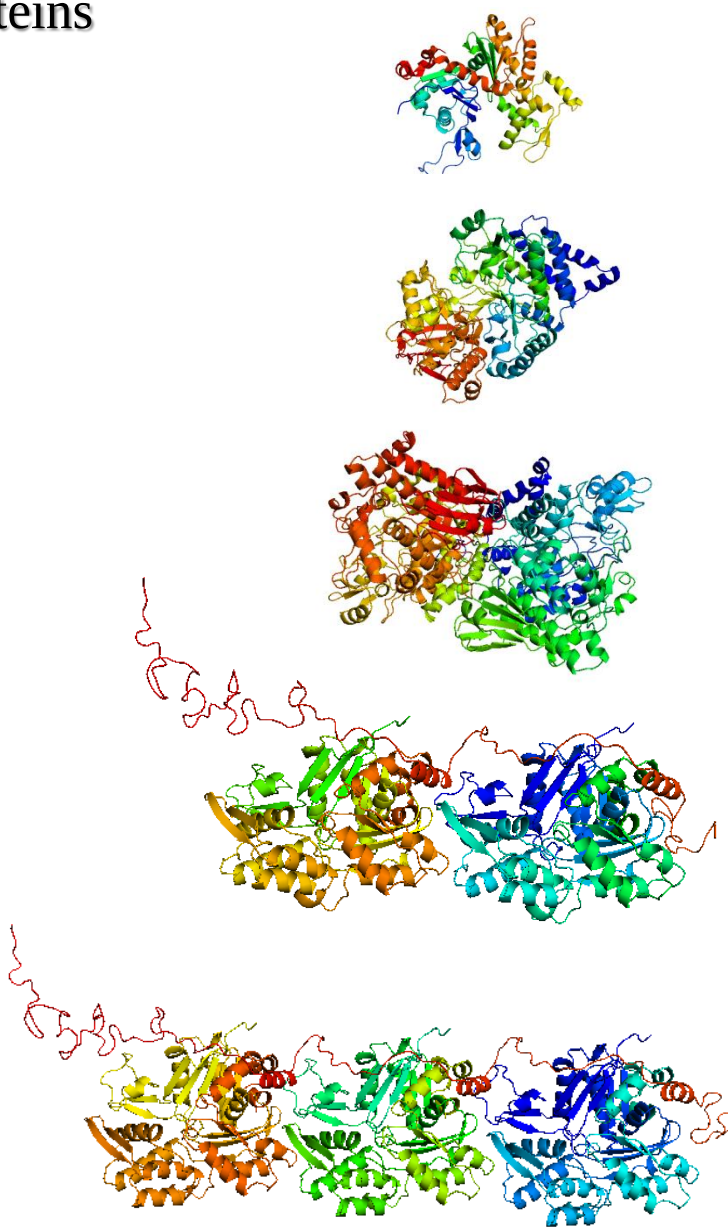
Folded particle : ***bell-shaped curve*** (asymptotic behaviour $I(Q) \sim Q^{-4}$)

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





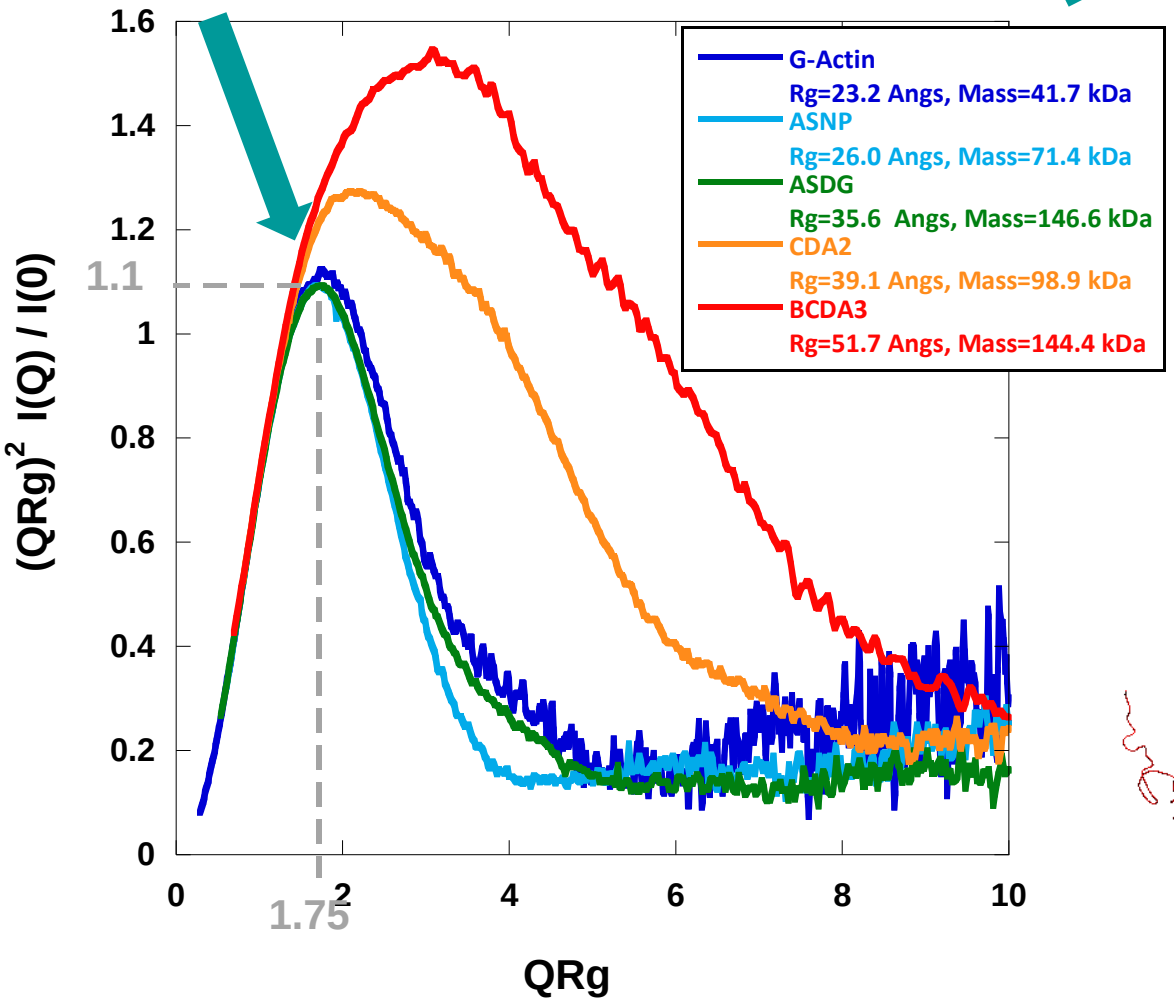
Folded proteins display a bell shape. Can we go further?

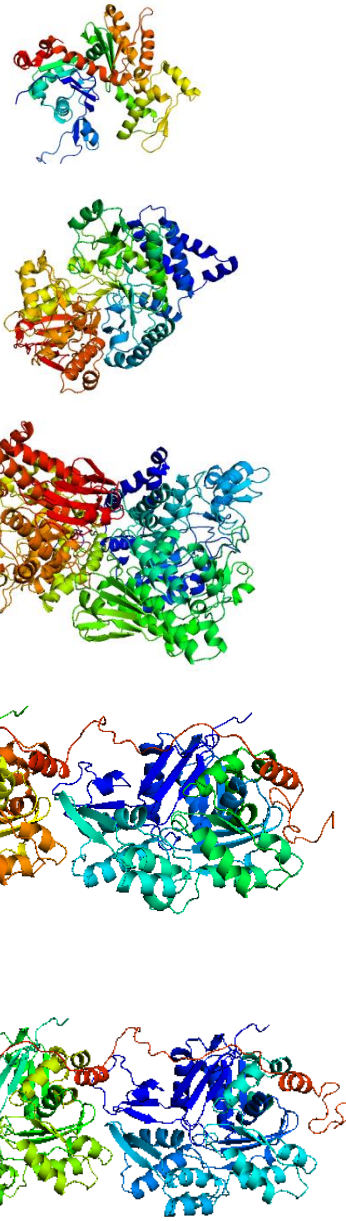
Dimensionless Kratky Plots of folded proteins

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

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

For globular structures, DLKPs fold into the same maximum

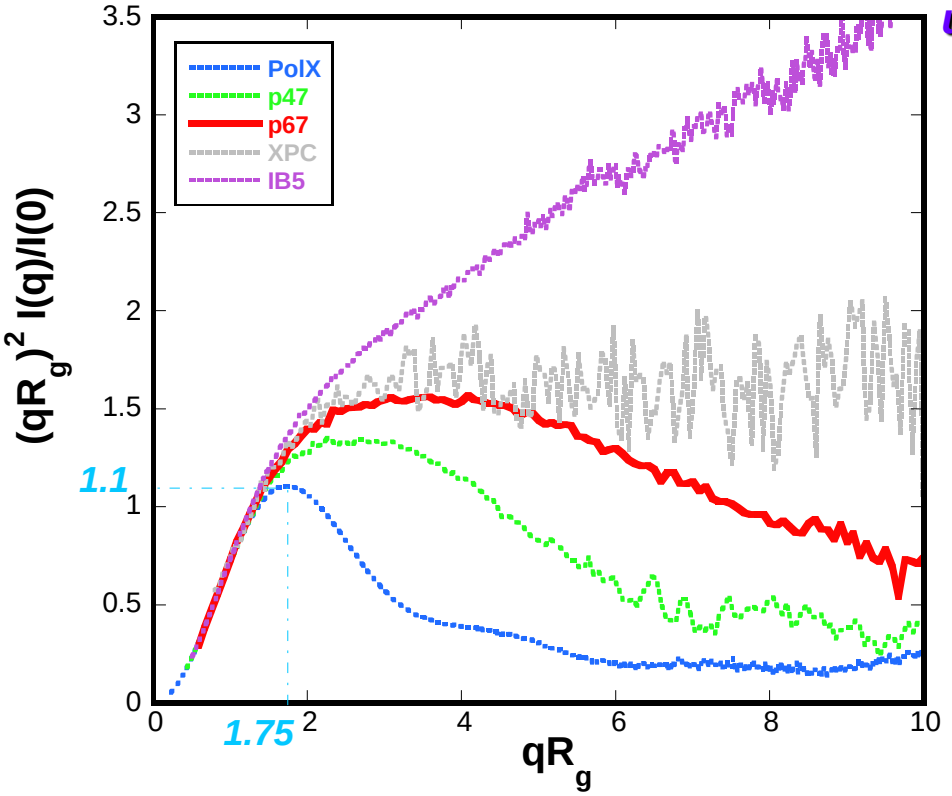




The maximum value on the dimensionless bell shape tells if the protein is globular.

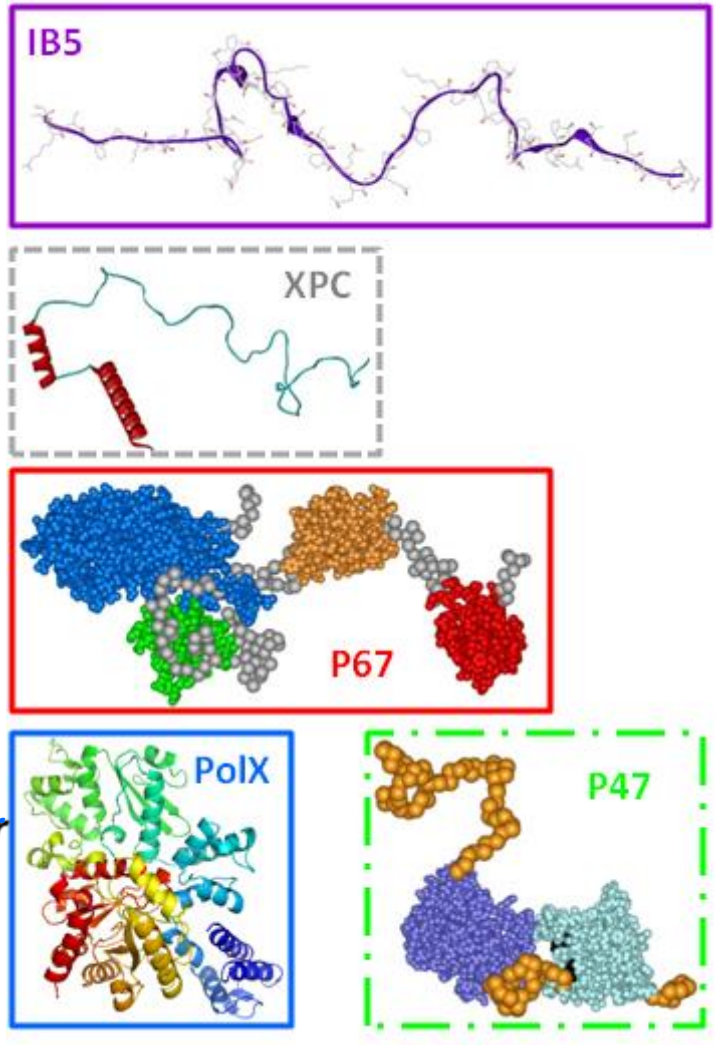
Dimensionless Kratky Plots of (partially) unfolded proteins

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.

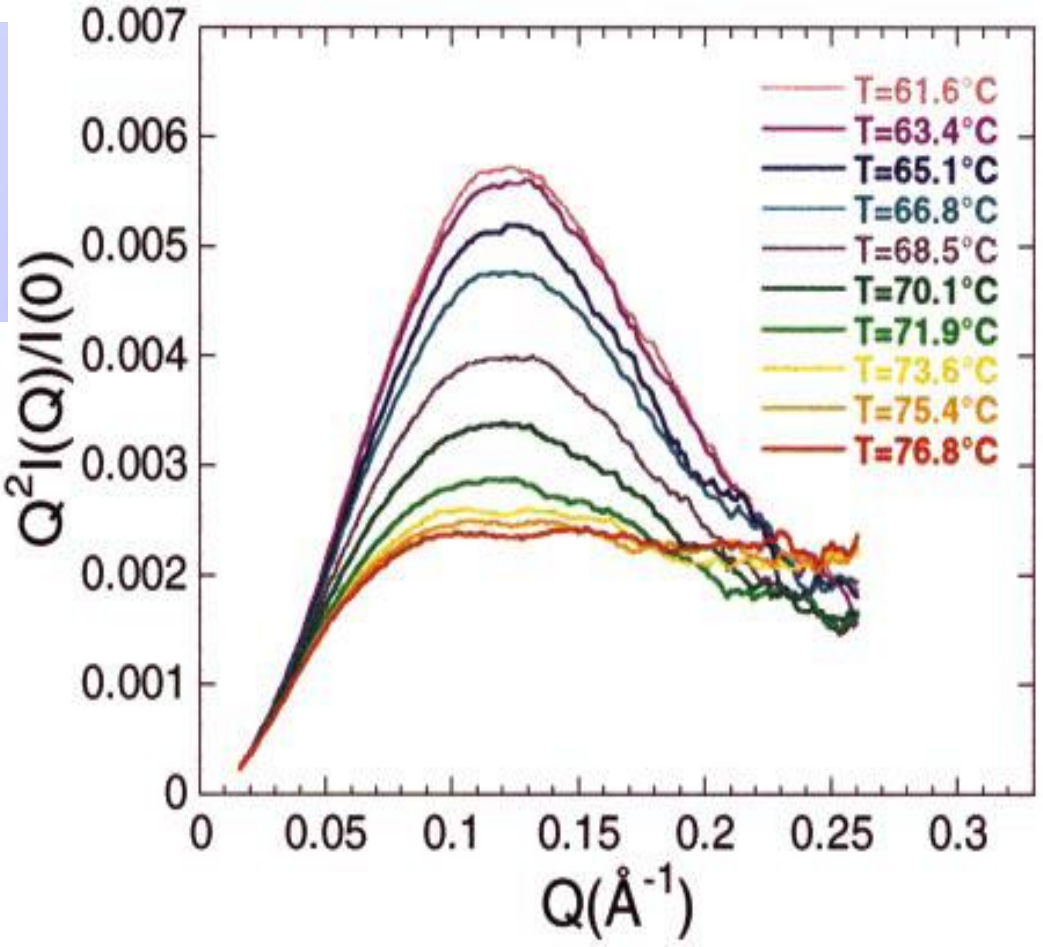
The curve increases at large Q as the structure extends.

Kratky Plot : NCS heat unfolding

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

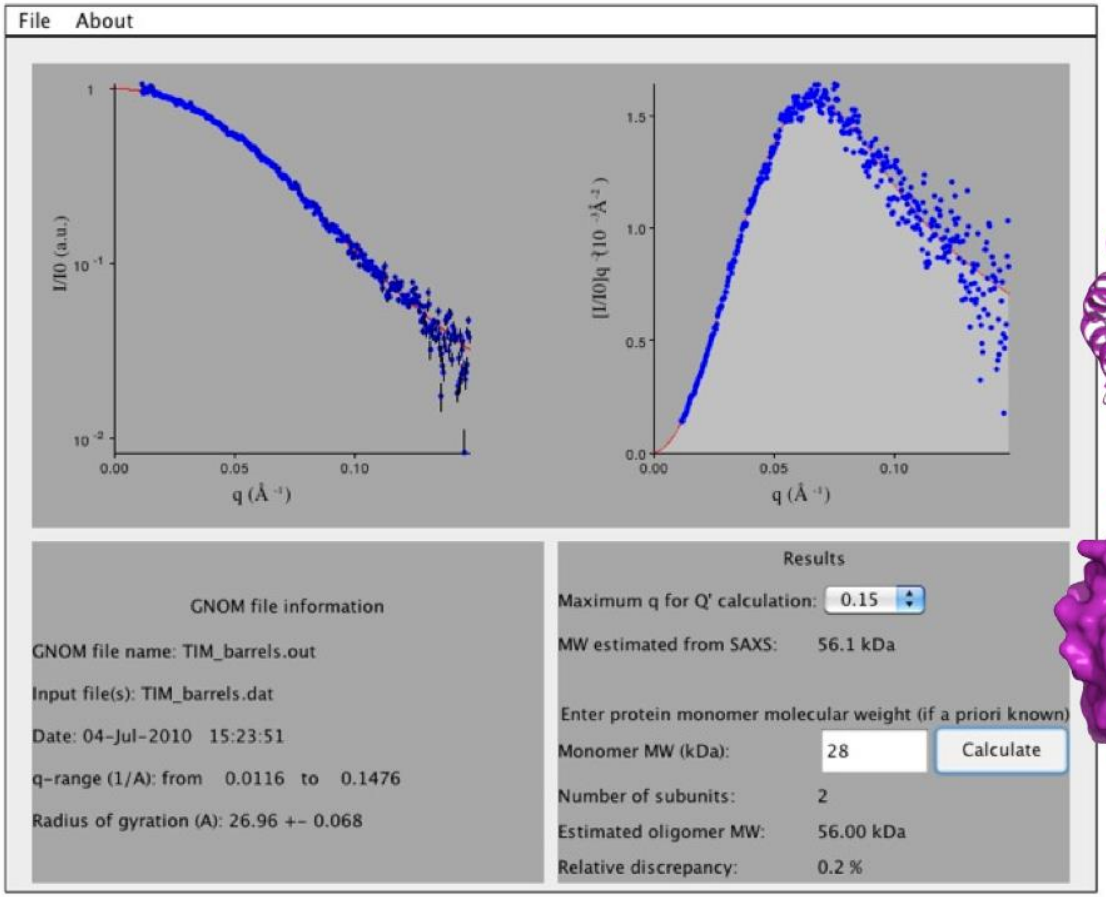


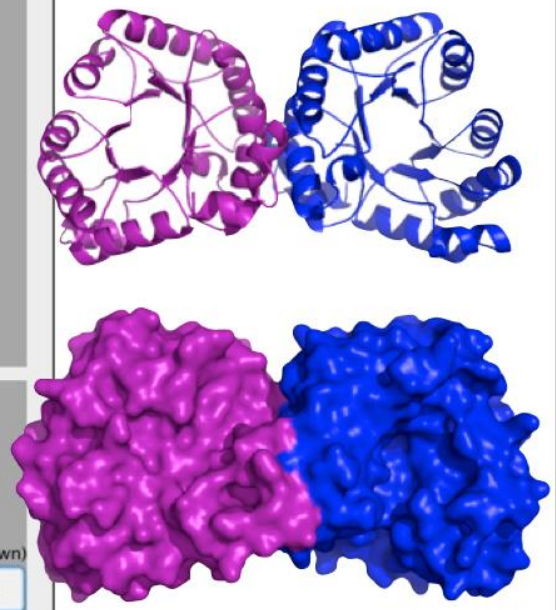
Pérez et al., *J. Mol. Biol.*(2001), 308, 721-743

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

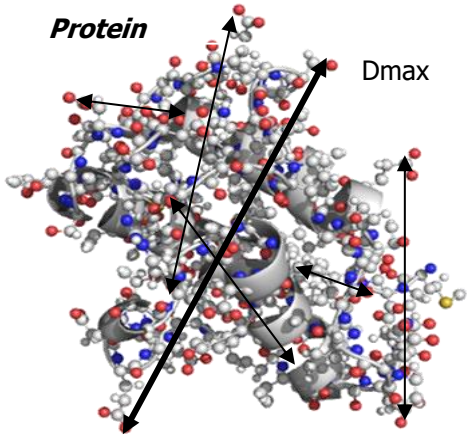
Rambo R. And Tainer J. (2013), Nature, **496**, 477-481.

Data Analysis

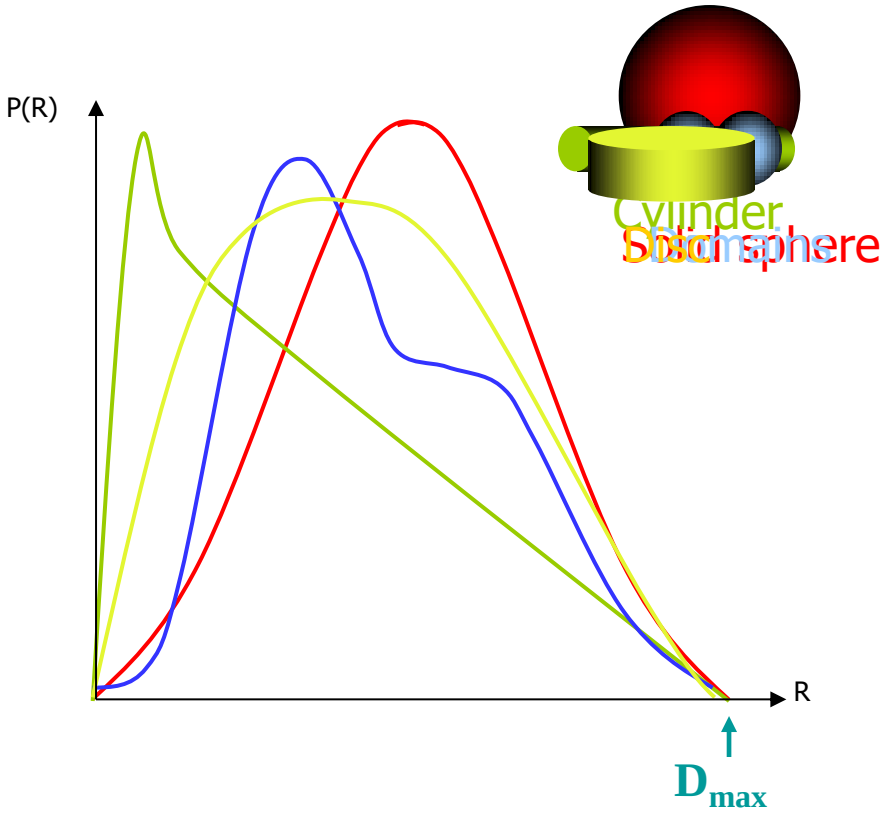
- Guinier Analysis
- Kratky plot : why is it so interesting ?
- « Real-space SAXS » : Distance correlation function P(r)

Distance Distribution Function p(r)

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_{max}



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

Relation between p(r) and I(q)

Intensity is the Fourier Transform of 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 :

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

Then :

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

And :

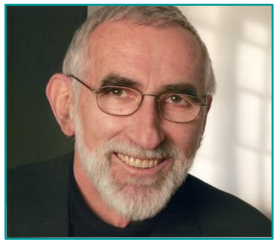
$$p(r) = \frac{r^2}{2\pi^2 \varphi r_e^{\,2}} \int_0^\infty q^2 I(q) \frac{\sin(qr)}{qr} dq$$

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 [Qmin,Qmax] 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
- p(r) is expressed over [0, D_{Max}] by a linear combination of orthogonal functions

$$p_{theoret}(r) = \sum_{1}^{M} c_n \varphi_n(r)$$

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

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

Dr. Dmitri Svergun
Hamburg, Germany

Svergun (1988) : program "GNOM"

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

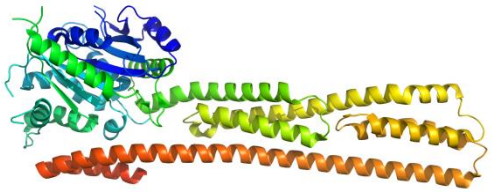
+ "Perceptual criteria" : smoothness, stability, absence of systematic deviations

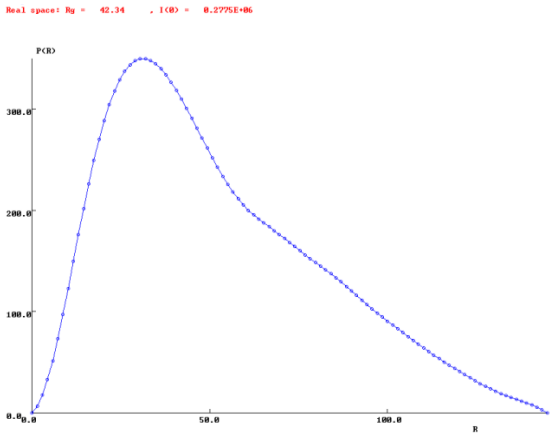
- Each criterium has a predefined weight
- The solution is given a score calculated by comparison with « ideal values »

Distance Distribution Function

Experimental examples

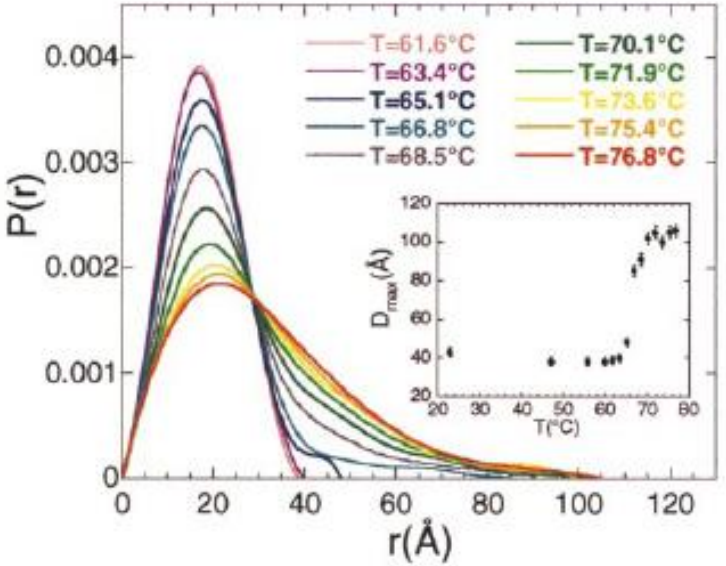
GBP1





Heat denaturation of Neocarzinostatin

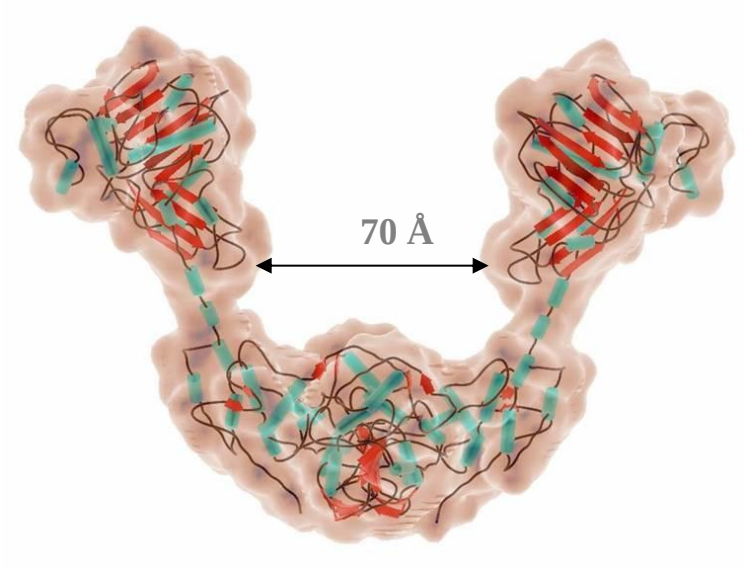
Pérez et al., J. Mol. Biol. (2001) 308, 721-743



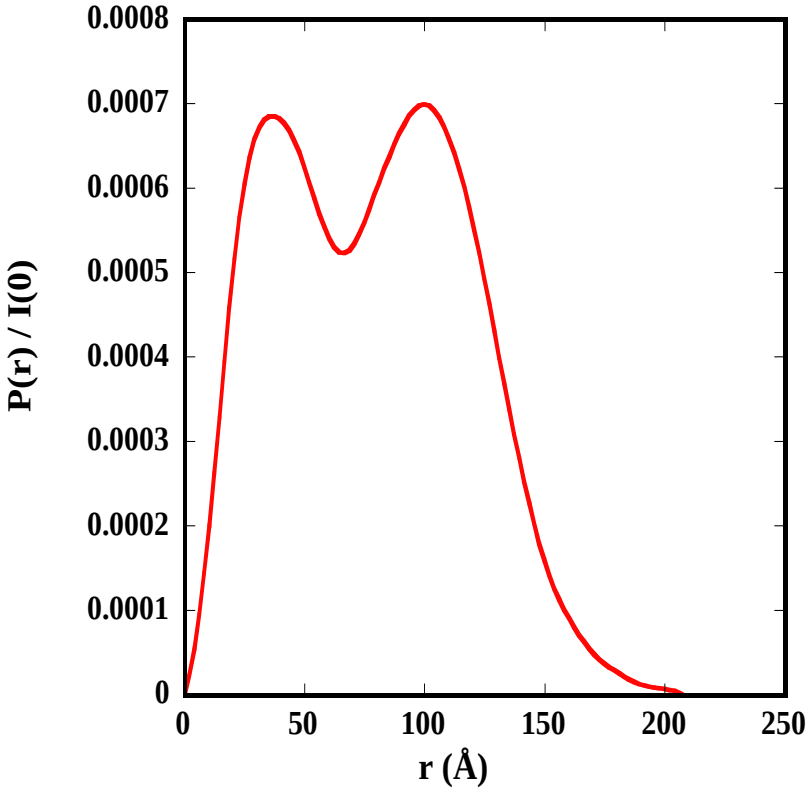
Distance Distribution Function

Experimental examples

Topoisomerase VI



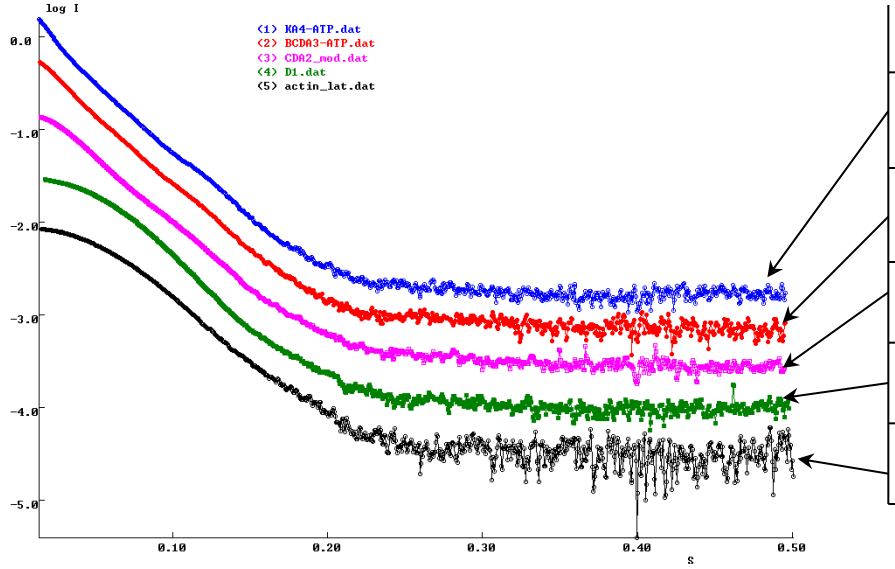
Bimodal distribution

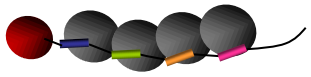


M. Graille et al., Structure (2008), *16*, 360-370.

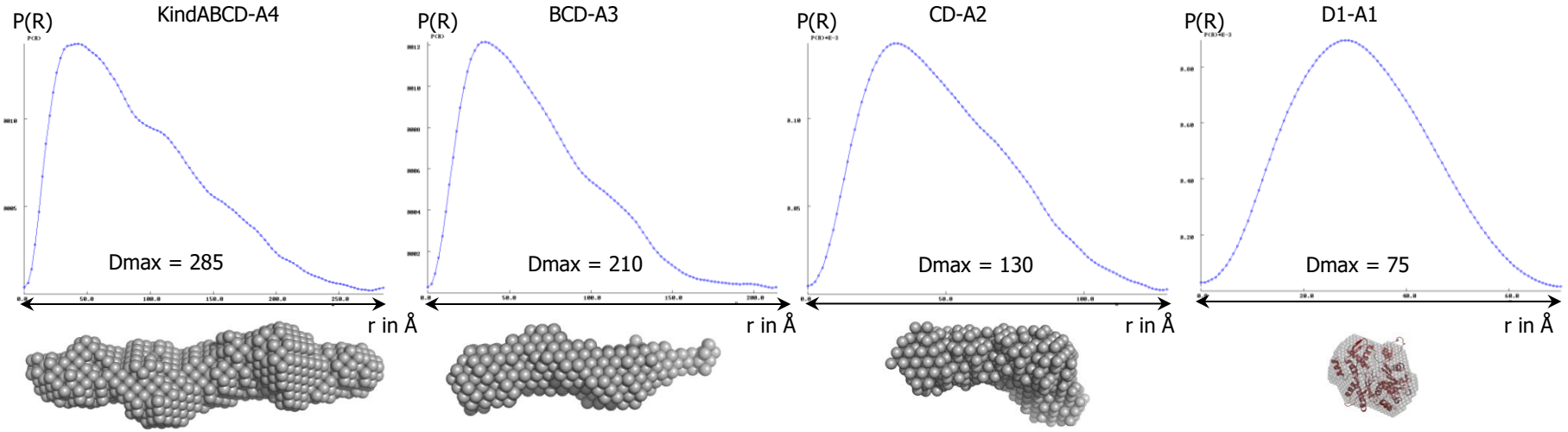
Distance Distribution Function

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



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

The radius of gyration and the intensity at the origin can be derived from p(r) using the following expressions :

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

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.

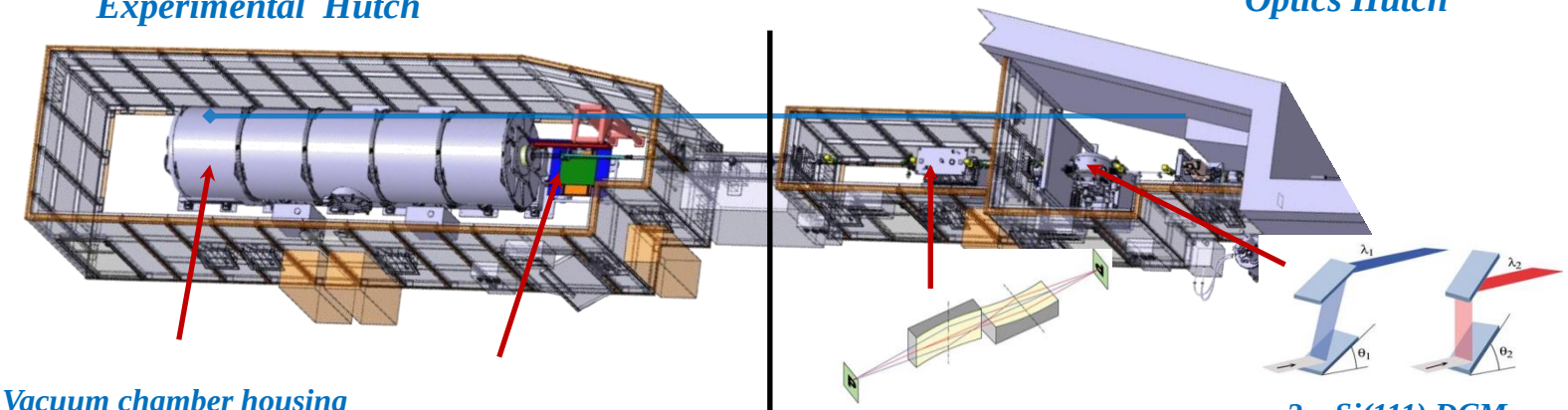
Comparison of estimates from Guinier analysis and from P(r) is a useful cross-check.

A FEW EXPERIMENTAL CONSIDERATIONS

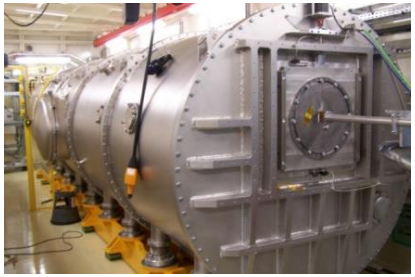
Schematics of beamline SWING

Experimental Hutch

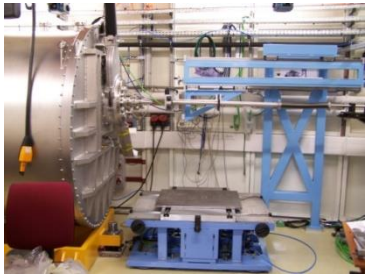
Optics Hutch



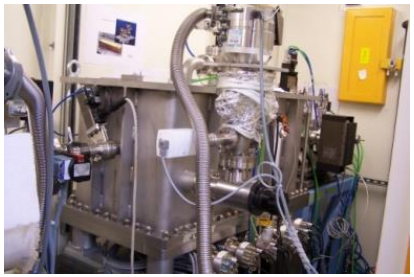
Vacuum chamber housing X-rays detectors



X-Z motorized table

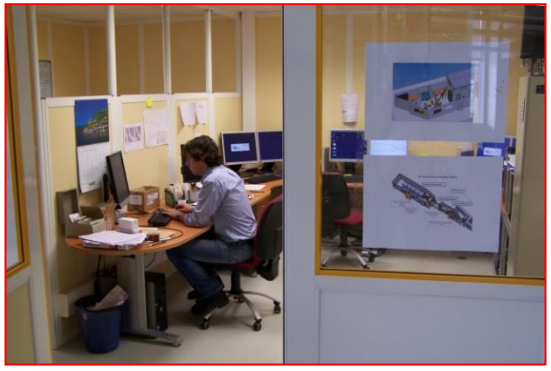


Curved mirros (KB config)



2 x Si(111) DCM 4.5 – 17 keV

Control room



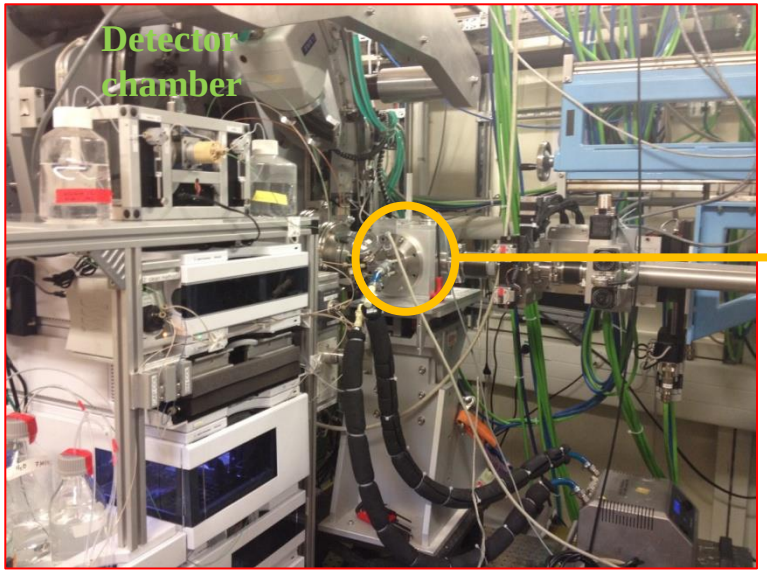
Full flux
5.10^{12} ph/s @ 12 keV

Beam size (FWHM)
400 (H) x 25-100 (V) μm²

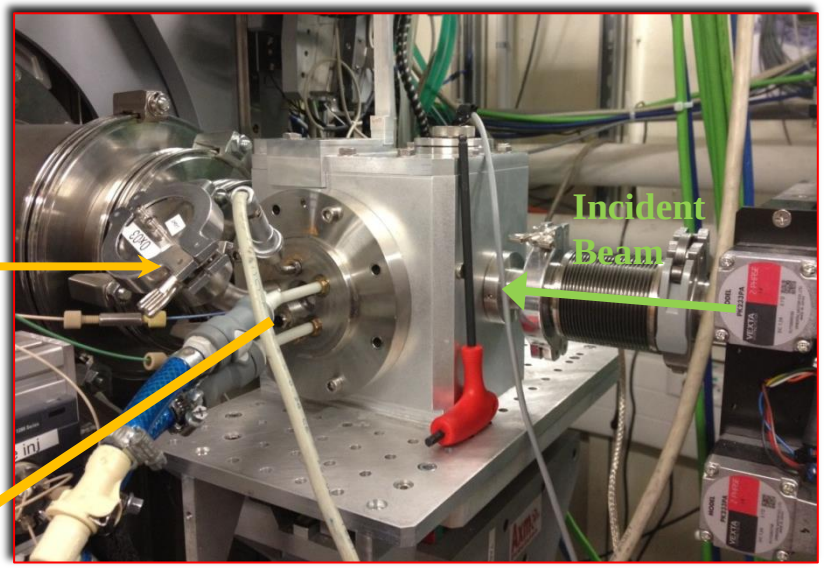
In-vacuum Undulator U20
gmin = 5.5 mm

Set-up for BioSAXS at Beamline SWING

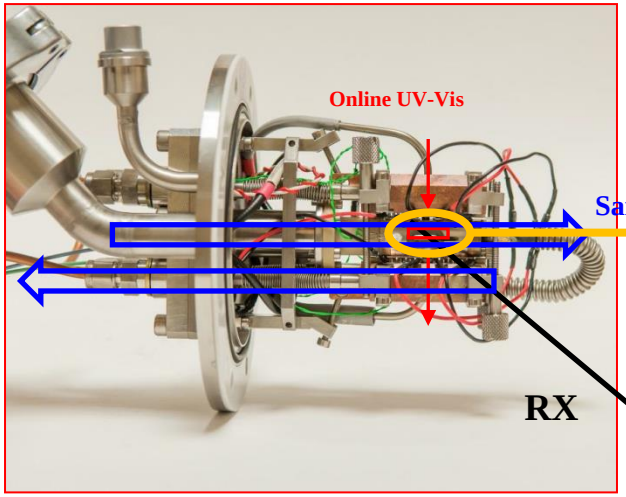
G. David and J. Pérez (2009), J. Appl. Cryst



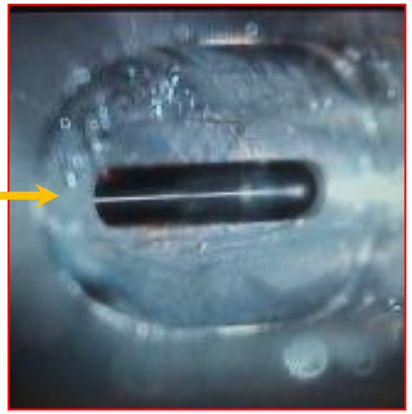
SEC-SAXS



BioSAXS Vacuum chamber



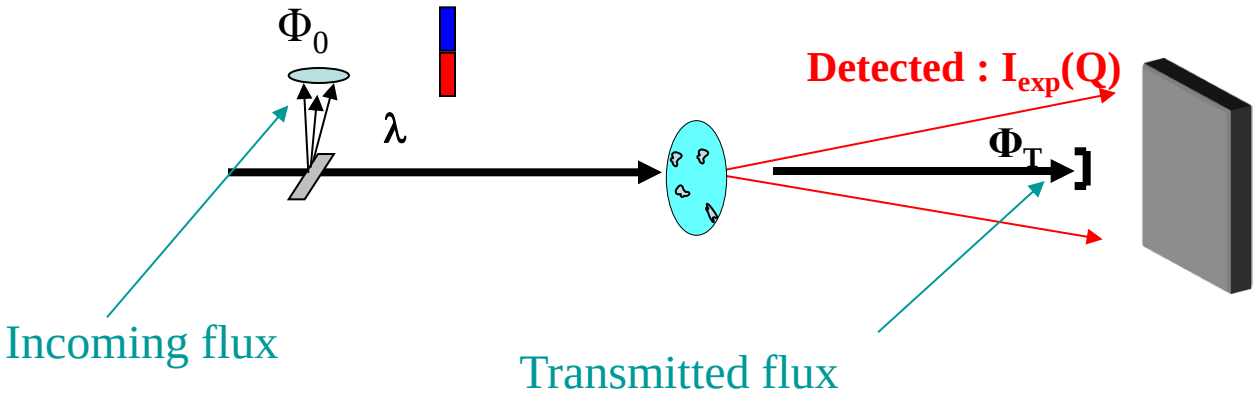
Details of the BioSAXS cell



Quartz capillary

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 %).

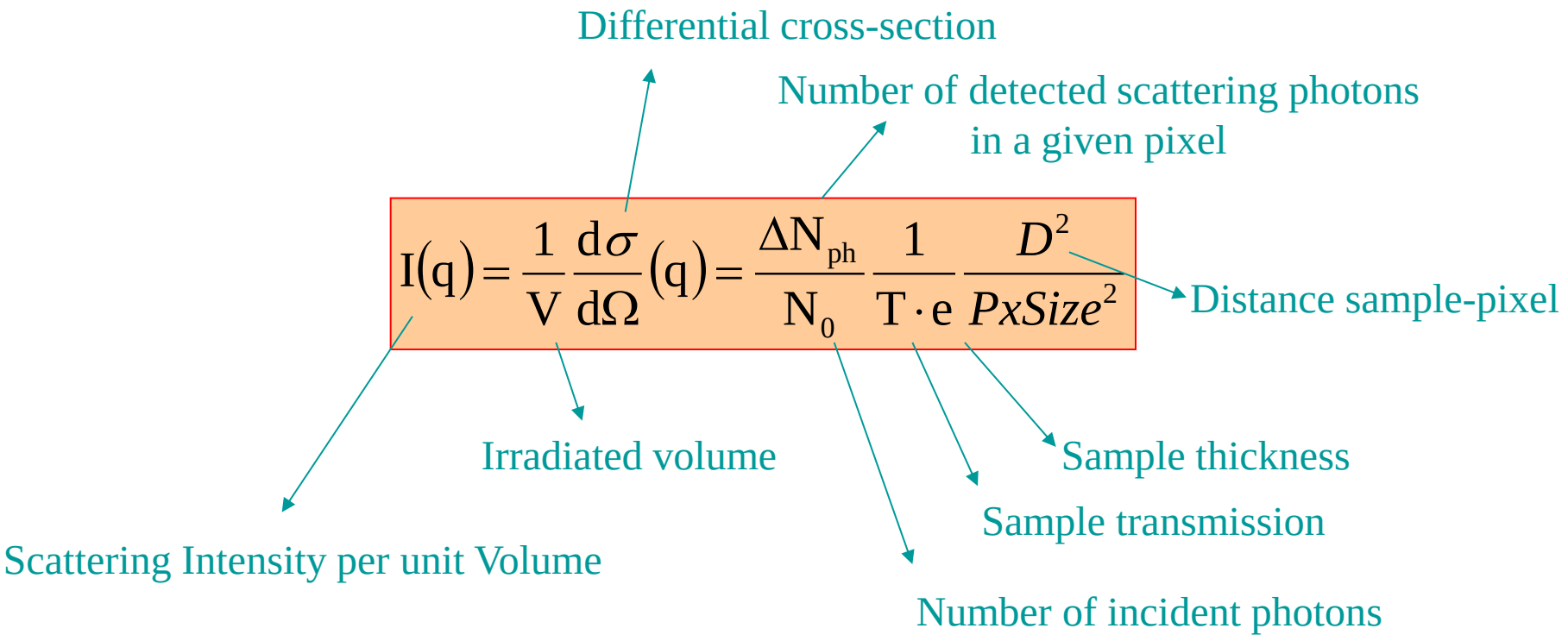


- Buffer
  - Buffer and protein samples must be measured in the same cell for correct subtraction of parasitic background arising from slits and holder walls.
  - The buffer in the buffer sample must be identical to that of the protein sample (dialysis, SEC, …).

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

Particles in solution

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

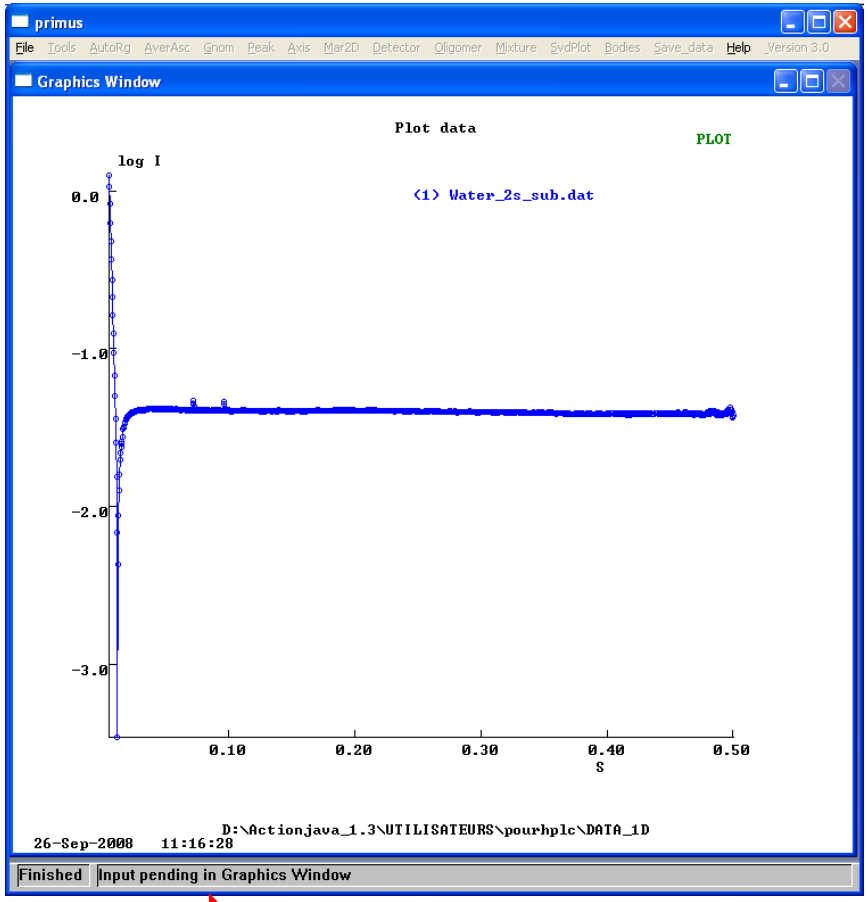


Calibration of the set-up using water scattering

SWING 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

$I_{H2O,theory} = 0.0163\ cm^{-1}$



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:

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

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

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

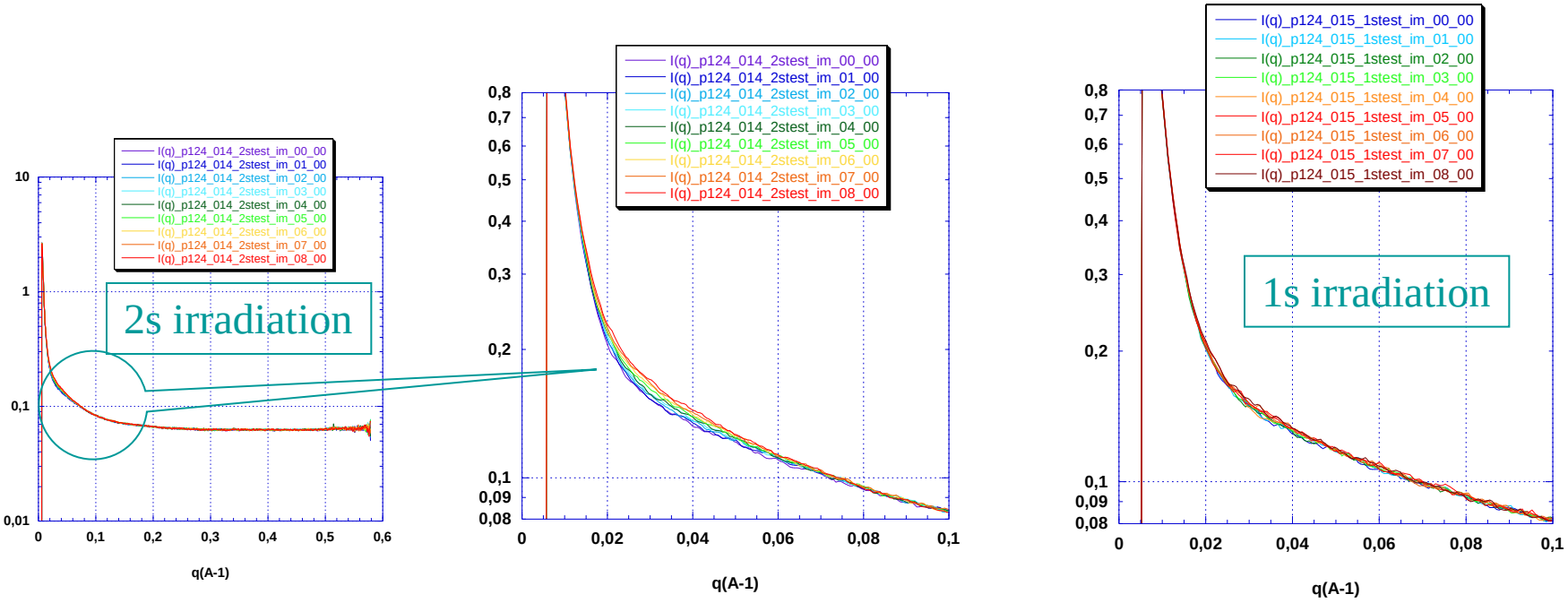
For any sample in that capillary : $I_{theory}(cm^{-1}) = I_{exp} / K_{exp} = I_{exp} / 2.56$

Protocol for data collection and treatment

Data collection

1st case : the solution is supposed to be monodisperse

- Test radiation damage (7µl) → determine frame irradiation time

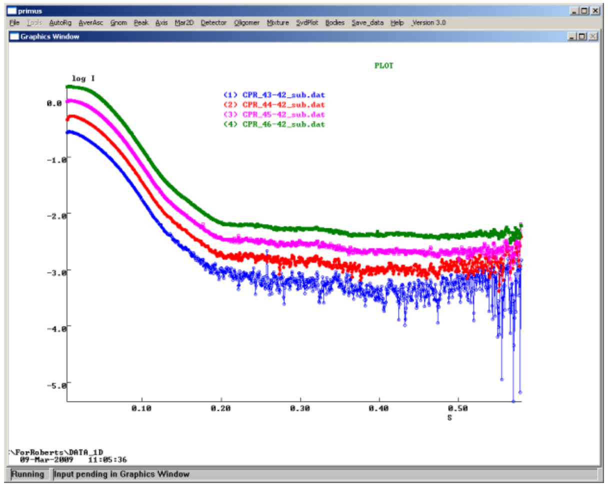


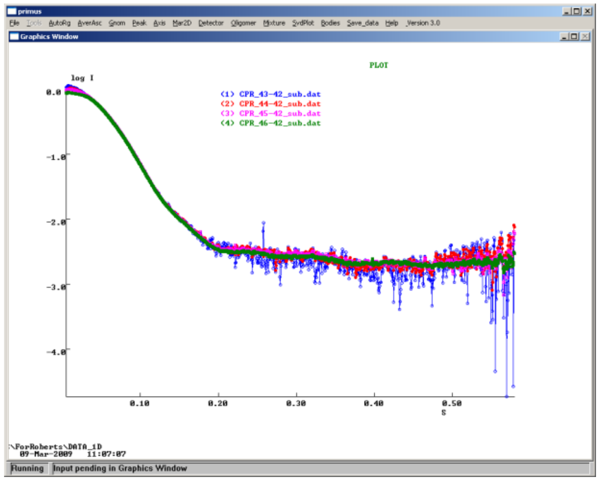
- Data collection on concentration series (25 µl) → take account of long range interactions

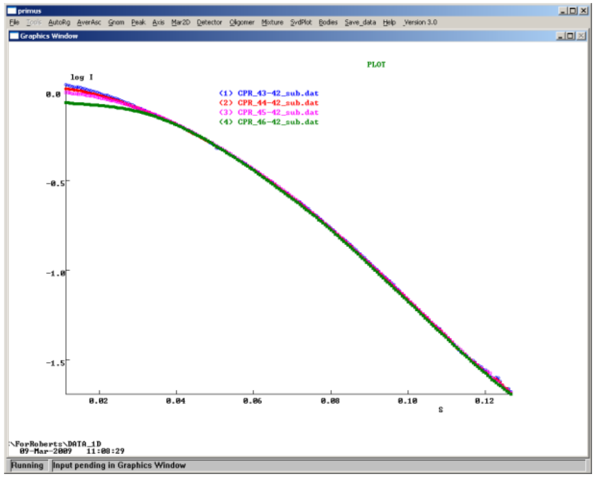
Protocol for data collection and treatment

Data treatment

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







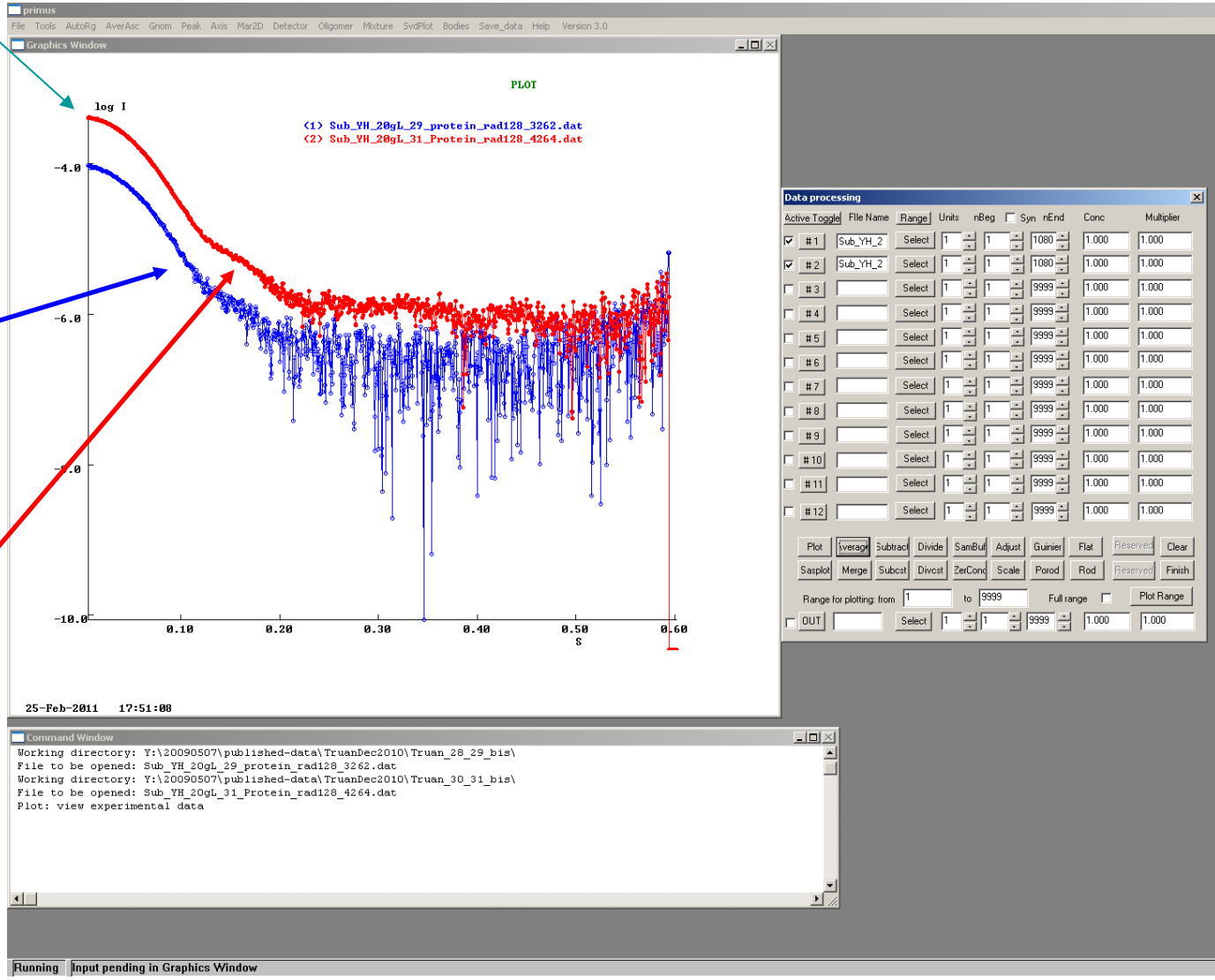
- **If necessary, merge** low c (low Q) and high c (high Q) curves
- **Compute p(r) →** should gently vanish at D_{max}

PRIMUS: combining data

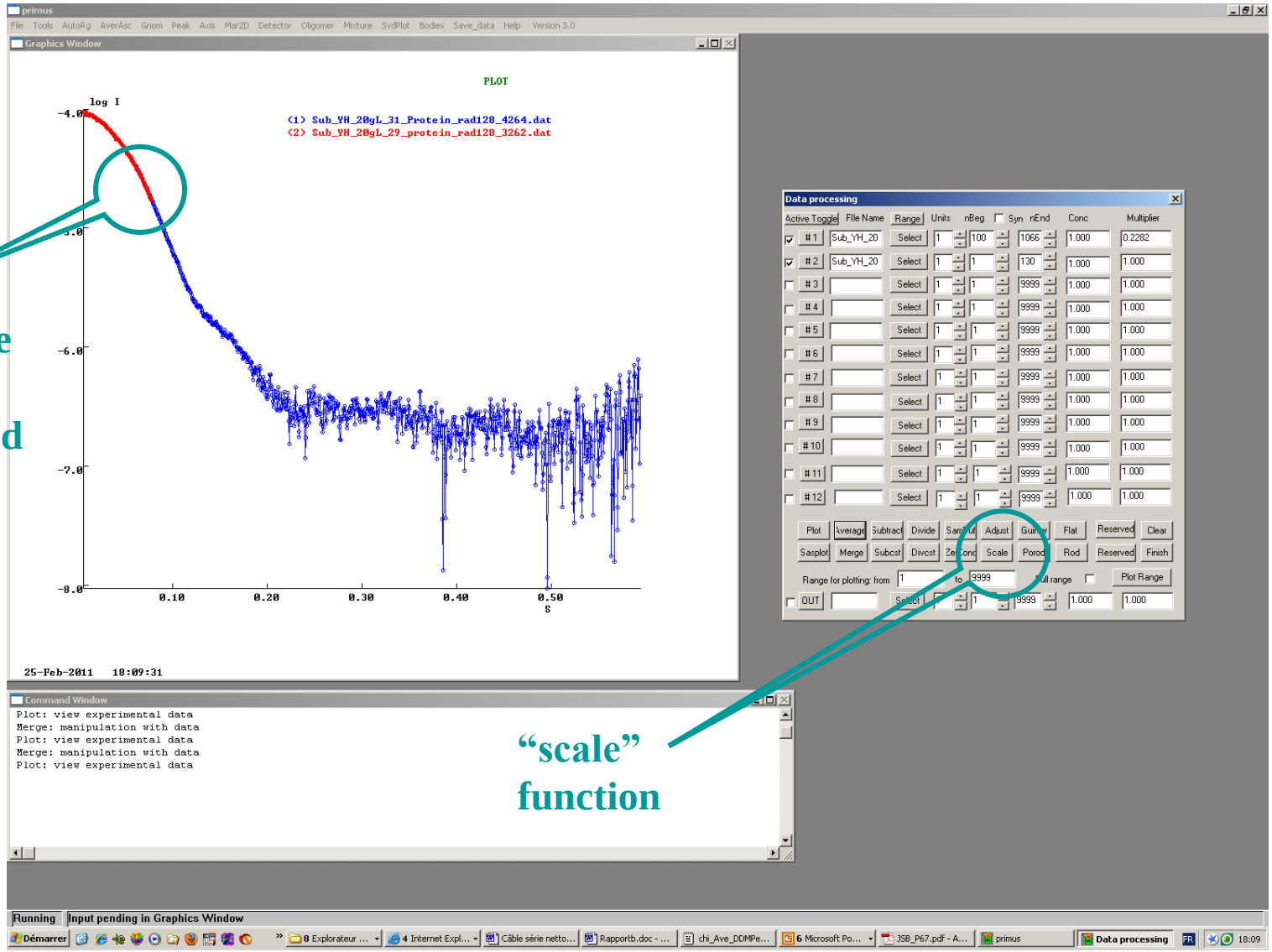
Here, slight repulsive interactions alter the concentrated curve at small angles

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



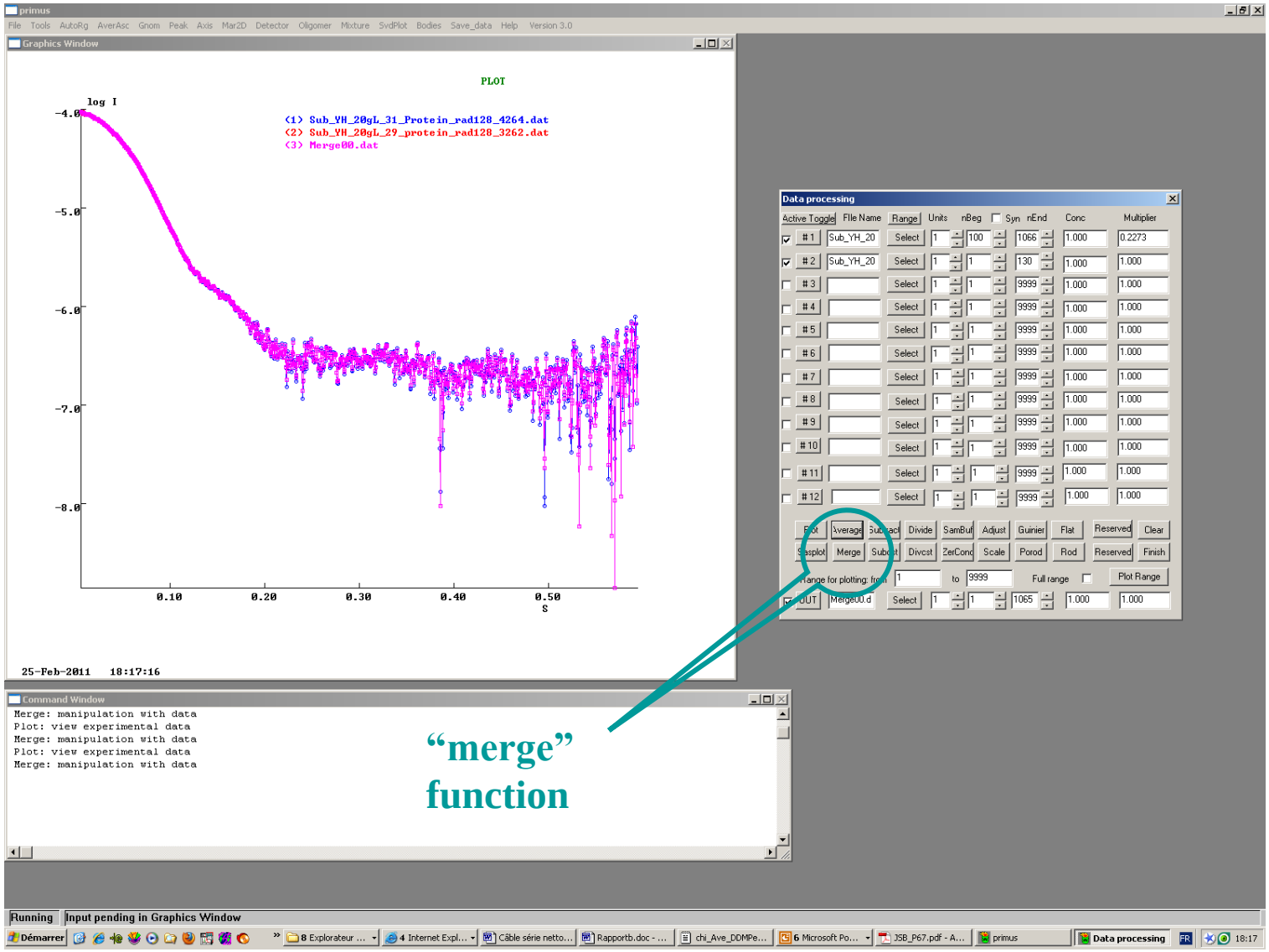
PRIMUS: merging data



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

“scale” function

PRIMUS: final merged curve

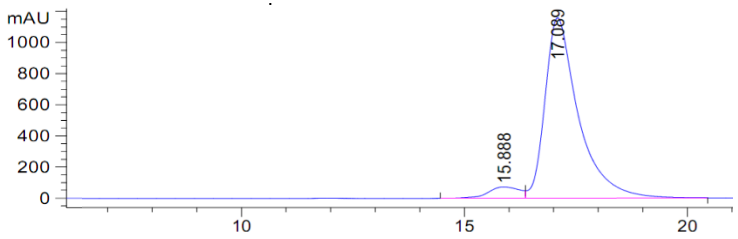


“merge” function

2nd case : the solution is a slow equilibrium or an unwanted mixture

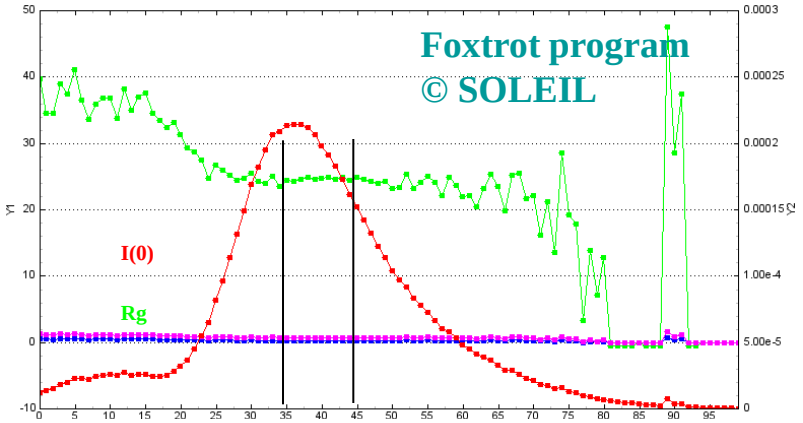
- Use on-line HPLC data collection (typ 50 µl)

ASNP elution profile, monitored by UV absoprtion at 280 nm



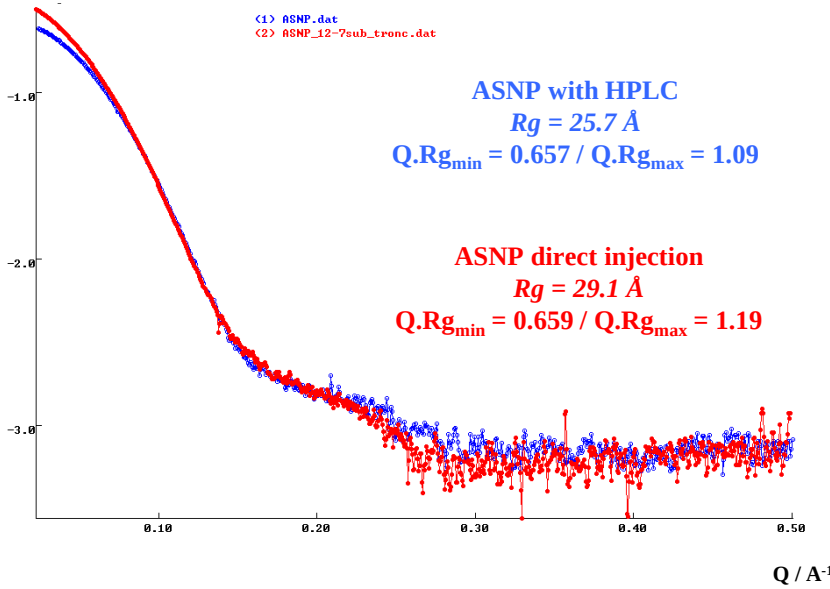
Elution time (minutes)

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

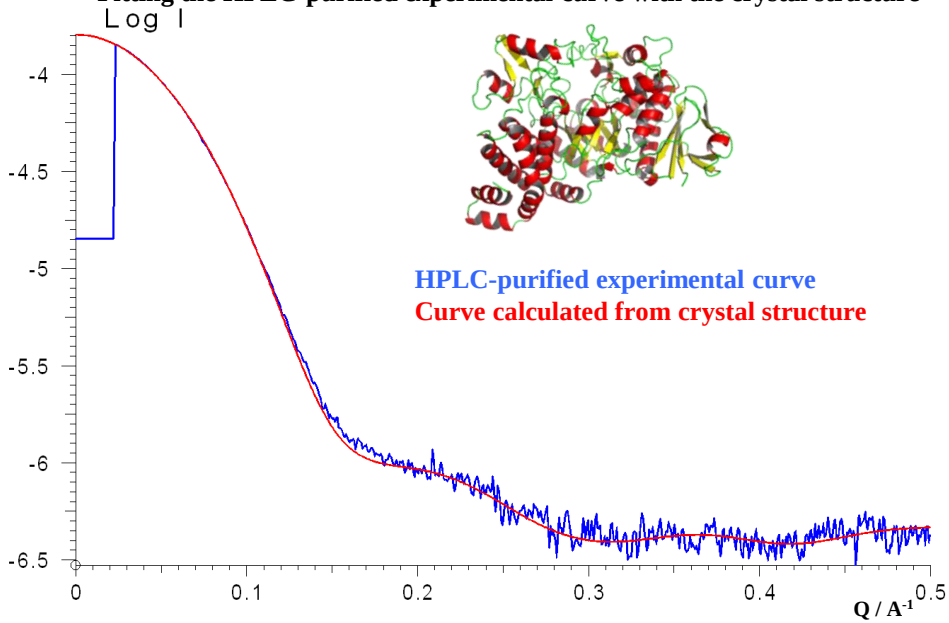


Frame number (~ time)

Comparison between HPLC-purified and Direct injection curves

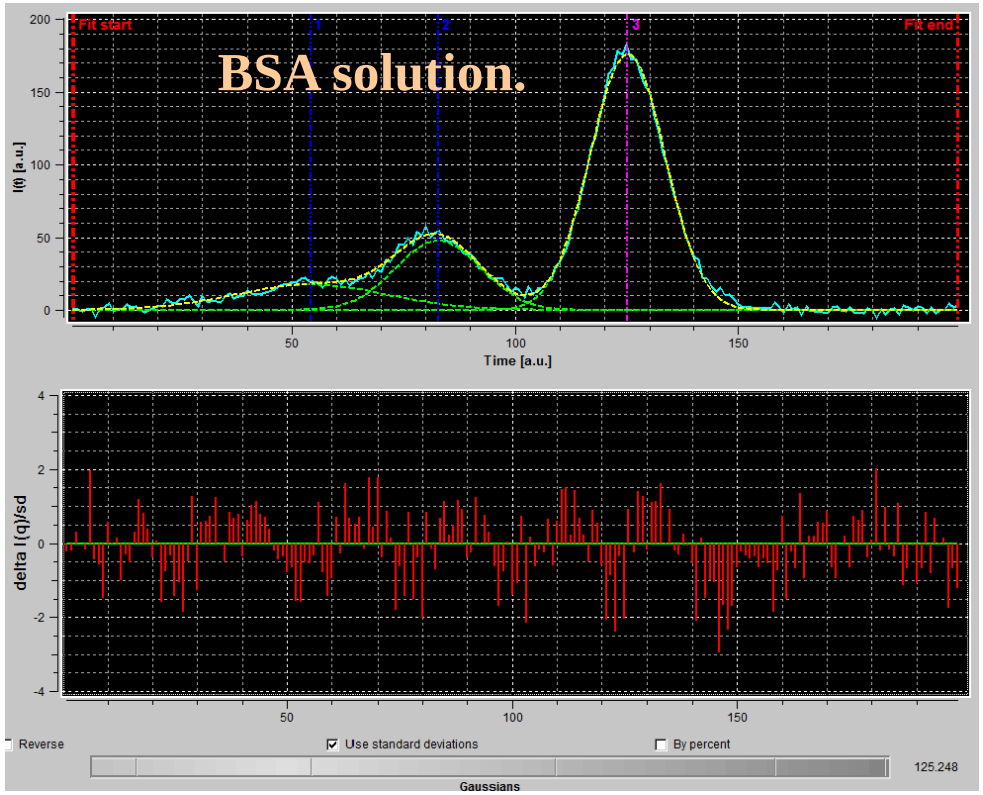


Fitting the HPLC-purified experimental curve with the crystal structure



SEC-SAXS analyzed with US-SOMO

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

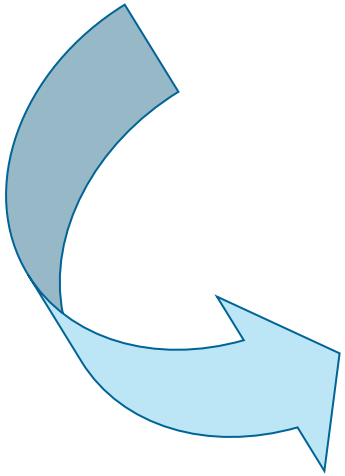


Following deconvolution, the scattering pattern $I_j(q)$ of each species j can be reconstructed.

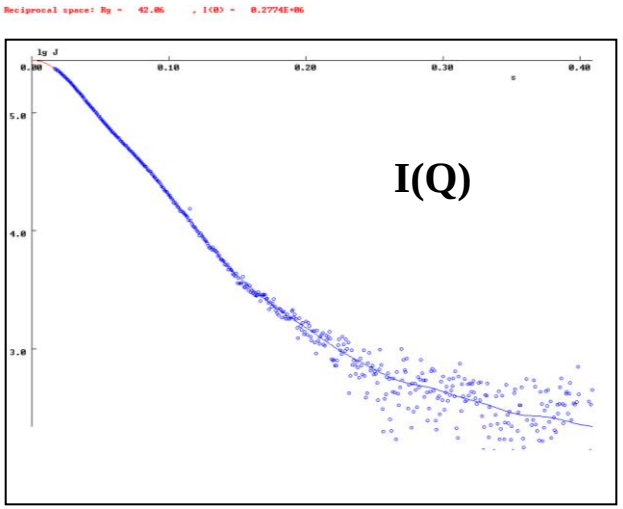
E. Brookes *et al.*, J. Appl. Cryst 2013, 46, 1823-33.

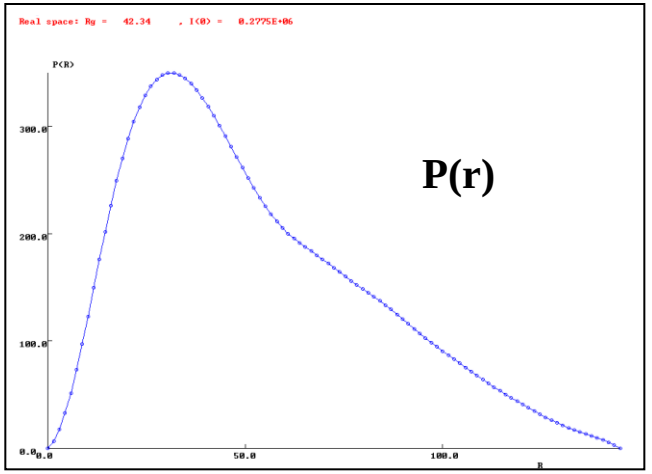
At this stage

We have gone from

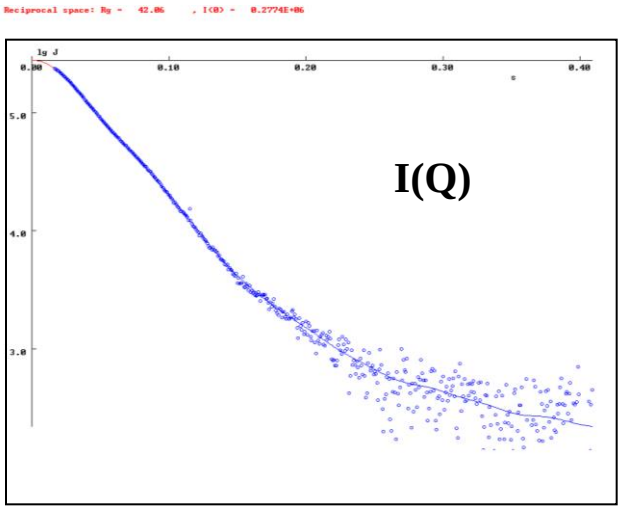


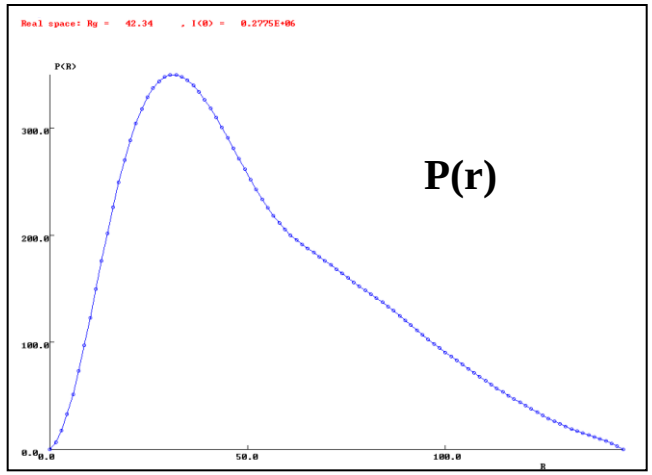
to



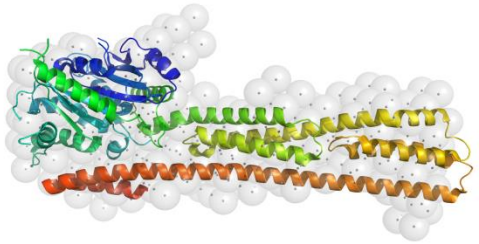


Now, we have to go from





to

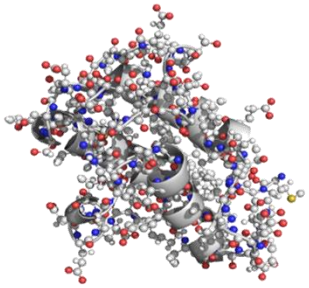


MODELLING

SAXS for 3D structure reconstitution

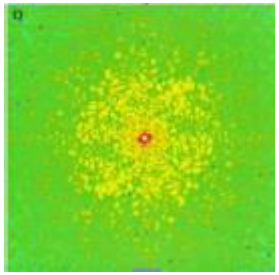
The 1D SAXS profile is the Fourier transform of the 3D structure.
But the inverse problem cannot be solved analytically, i.e., no "inverse computation" can be used to yield 3D position coordinates from scattering data.

Real space 3D Molecule



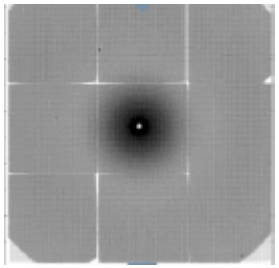
Fixed orientation → **Phase lost**

Reciprocal space 2D anisotropic image



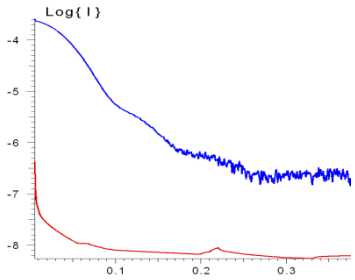
Averaged orientation → **Orientation lost**

Reciprocal space 2D isotropic image



Radial averaging

1D profile reciprocal space



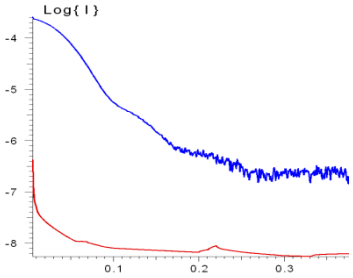
How to reconstruct the 3D structure from the 1D SAXS profile ?

Bear in mind !

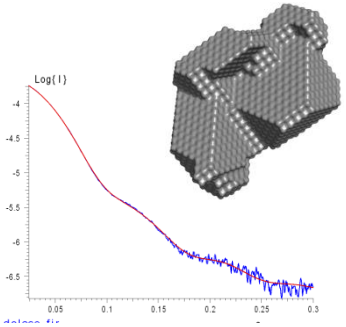
One 3D structure → One SAXS curve
BUT
One SAXS curve → Many 3D structures, all compatible with the same curve
Additional constraints are always needed

SAXS data analysis, available programs

1) Nothing known (except the curve)

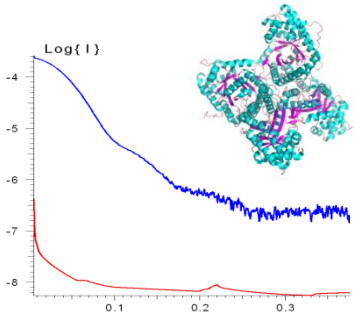


Low resolution model

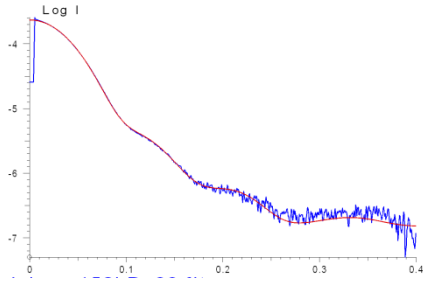


DAMMIN
DAMMIF
GASBOR
MONSA
DENFERT

2) Theorical model or complete atomic structure available

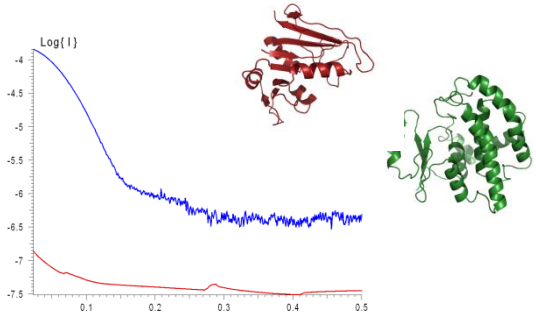


Validation/identification in solution



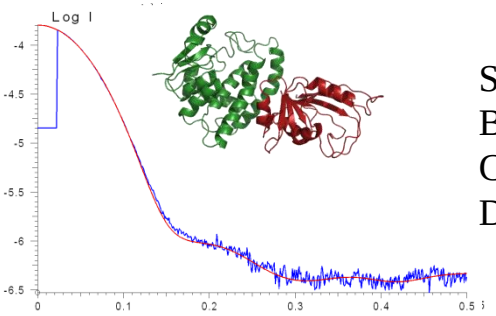
CRYSOL
FOXS

3) Structures of subunits available



Rigid body modeling of the complex and

molecular modeling of the missing part



SASREF
BUNCH
CORAL
DADIMODO

A word of caution

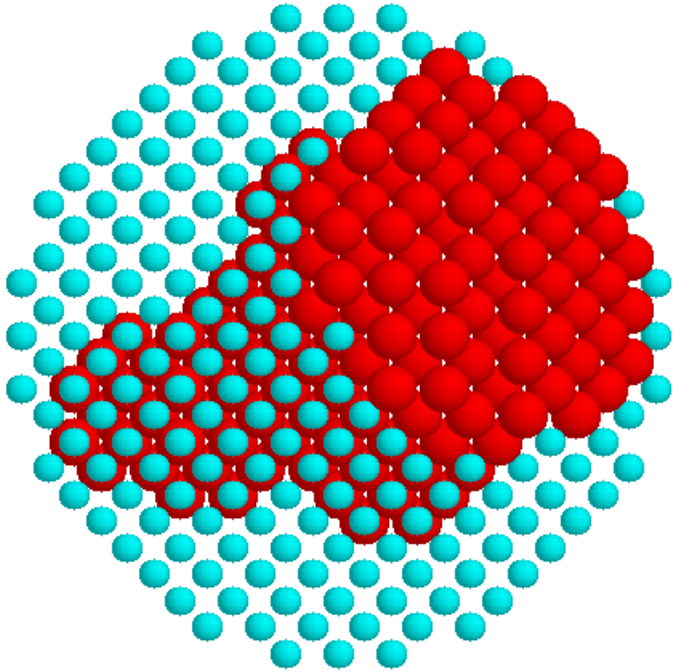
- $s = 2\sin\theta/\lambda$ **modulus of the scattering vector**
- $Q = 2\pi s = 4\pi\sin\theta/\lambda$ **momentum transfer**
- **But in his programs : D. Svergun uses**

  $s = 4\pi\sin\theta/\lambda$

Ab initio shape modelling : nothing is known but the curve

Ab initio shape modelling using a network of beads

Initial volume : sphere diameter D_{max}



Svergun, D.I. (1999) *Biophys. J.* **76**, 2879-2886

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

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

Chacón, P. *et al.* (1998) *Biophys. J.* **74,** 2760-2775.

Walter, D., Cohen, F.E. & Doniach, S. (1999), J. Appl. Cryst.,33, 350-363

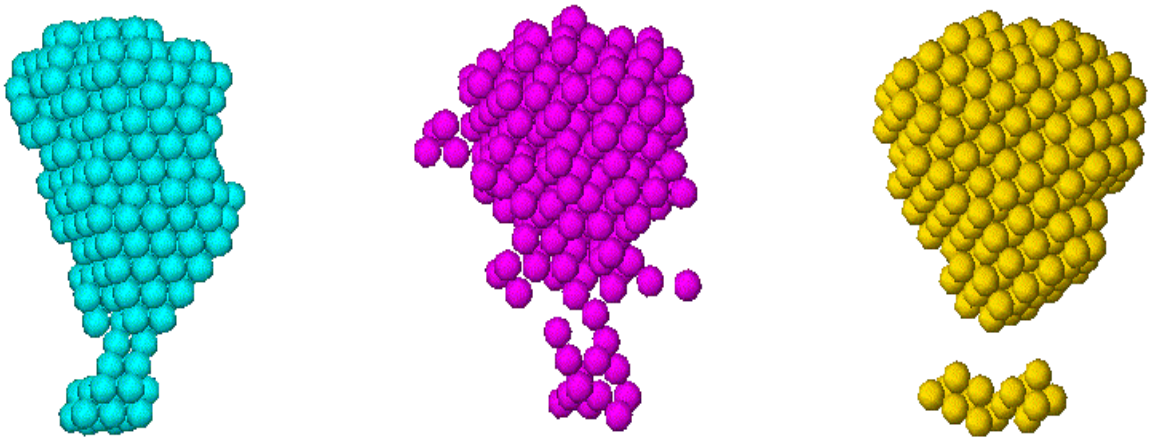
3D shape reconstructions from SAXS data with DAMMIN

- 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

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

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.



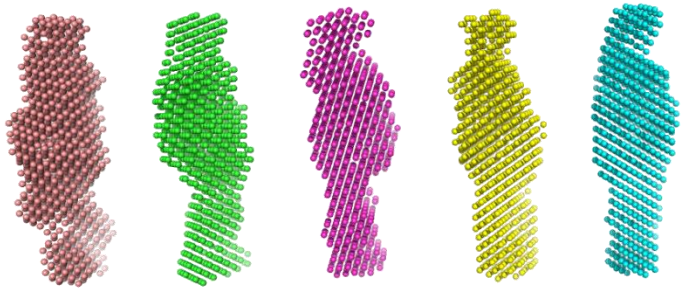
compact

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



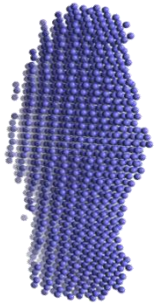
Shp1 Shp2 Shp3 Shp4 Shp5

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

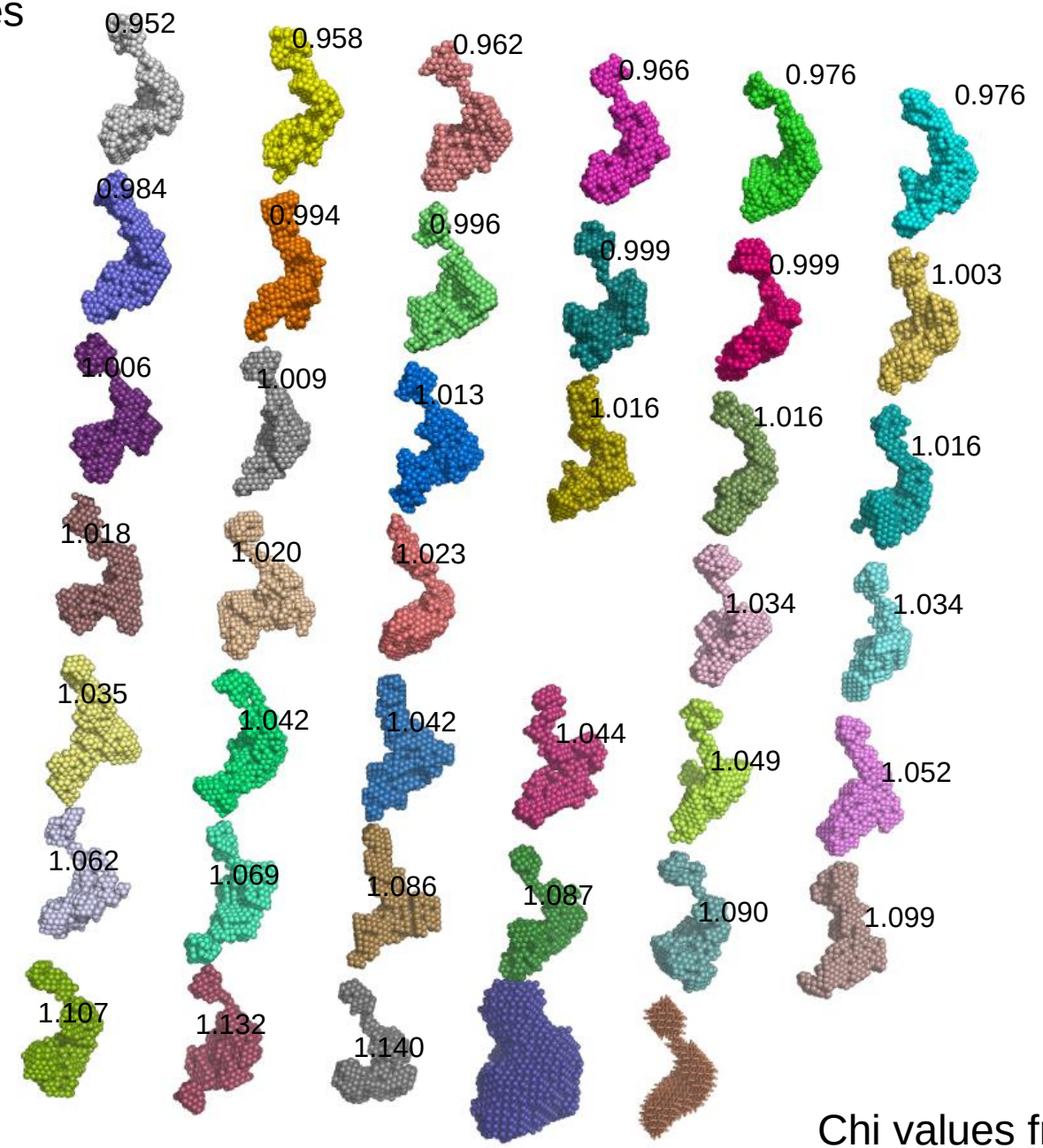
Damfilt (intersection)



Damaver (all superimposed)

- Models are conserved if its NSD < Mean of NSD + 2*standart 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 $I_{damfilt}(q) \neq I_{exp}(q)$

NSD values

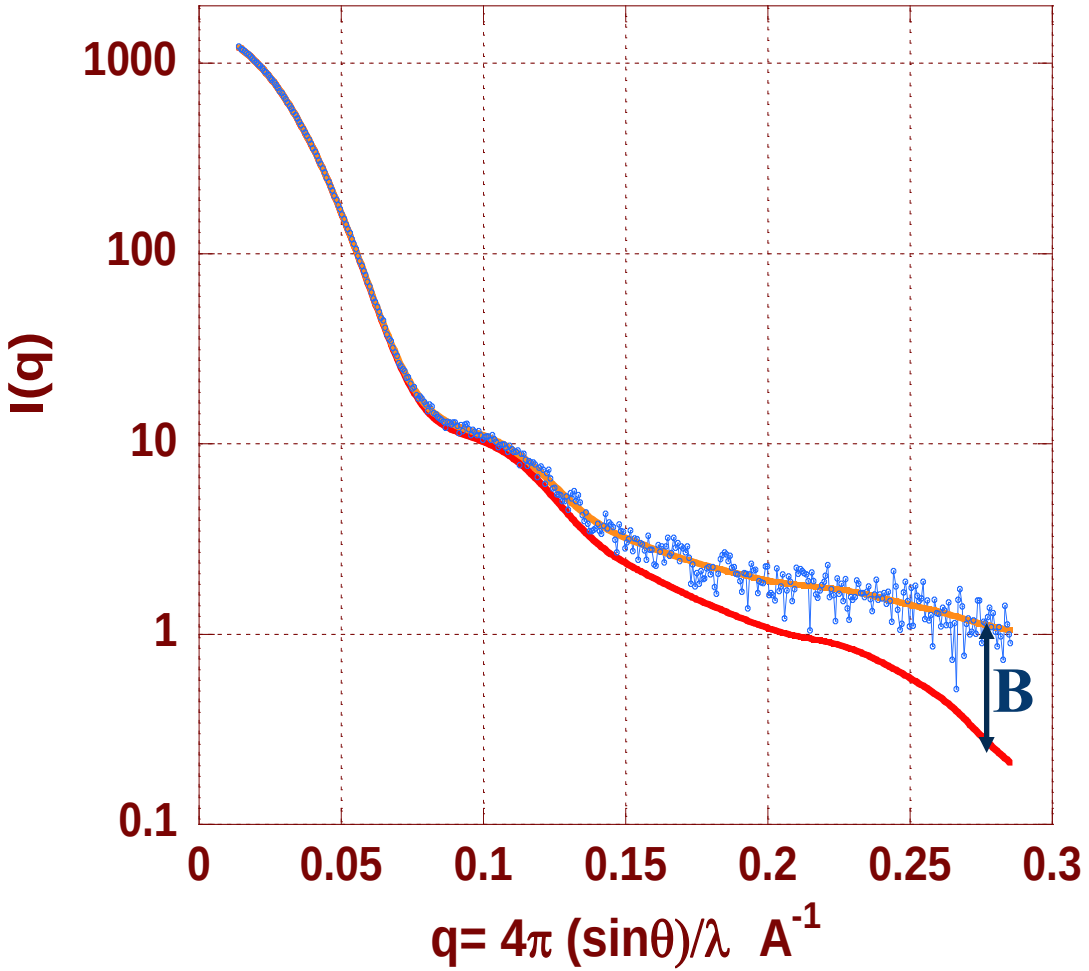


Chi values from 1.704 to 1.942

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.

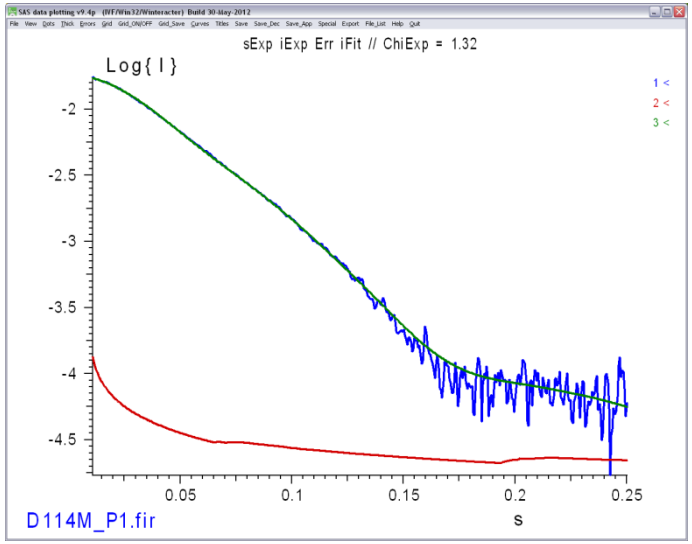


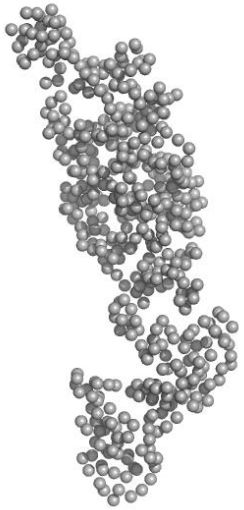
Ab initio model accounting for high resolution data

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

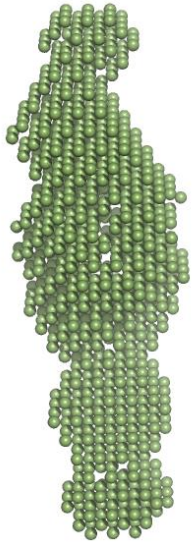
GASBOR : a protein comprising N residues is represented by an ensemble of N spheres centered at the Cα positions, the whole q-range can be used.

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

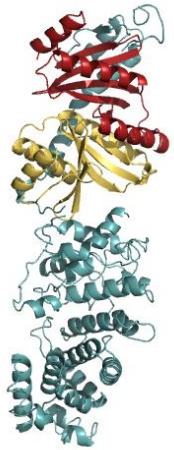




GASBOR beads model



DAMMIF shape



High resolution structure

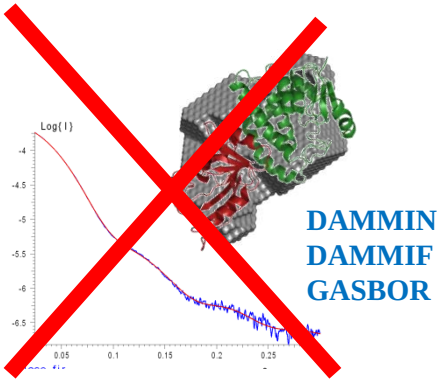
D. Svergun et al.(2001), Biophys. J., 80, 2946-2953.

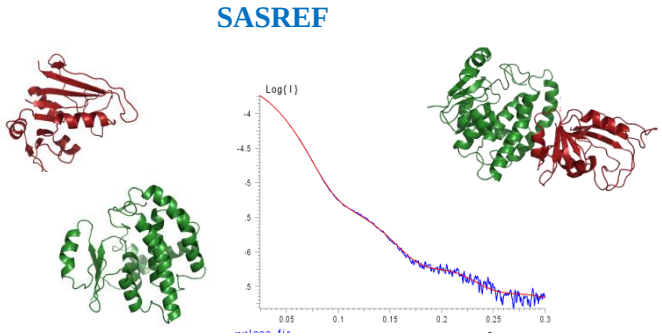
A word of caution: what NOT to do

- **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.**

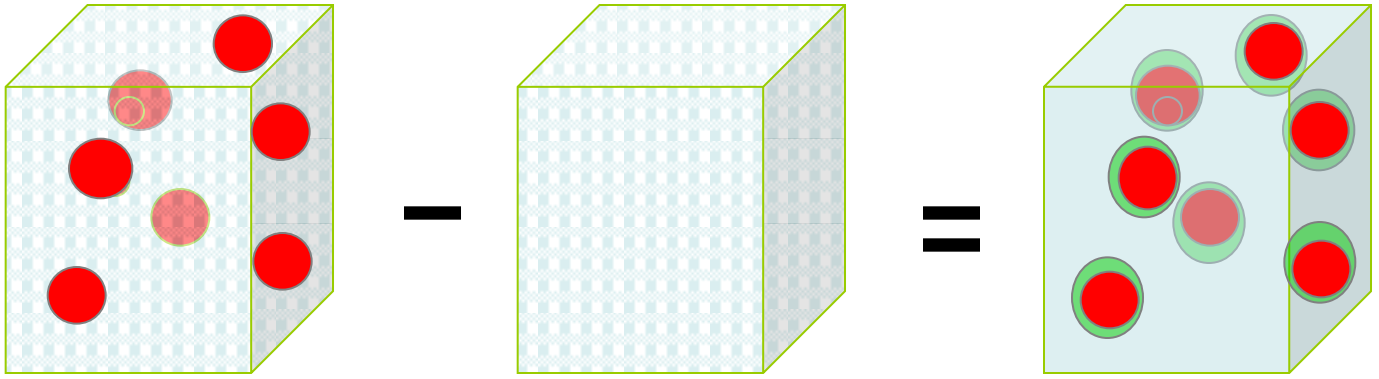
- **One must try and refine the position of domains vs SAXS data.**





From a atomic structure to a solution scattering pattern : program CRYSOL

Solvent scattering and contrast



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

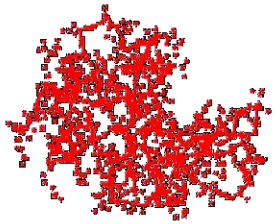
The bound solvent density differs from that of the bulk.

Bulk water density = 0.334 $e^-/Å^3$

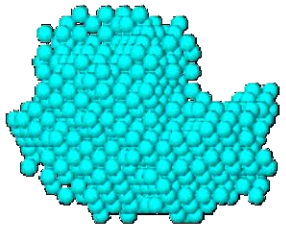
Hydration layer density ~ 5-15 % higher

Scattering from a macromolecule in solution

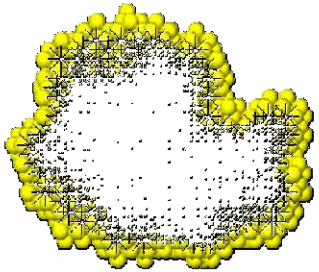
$$I(s) = \left\langle |A(\mathbf{s})|^2 \right\rangle_\Omega = \left\langle |A_a(\mathbf{s}) - \rho_s A_s(\mathbf{s}) + \delta\rho_b A_b(\mathbf{s})|^2 \right\rangle_\Omega$$



- $A_a(s)$: 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

CRYSOL ***(X-rays):*** **Svergun et al. (1995). *J. Appl. Cryst.* 28, 768**

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 developped in a series of Bessel functions and Spherical harmonics**

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

The experimental scattering curves are then fitted using only 3 parameters:

- **the general scale of $I_{calc}(Q)$**
- **the total excluded volume V, which is equivalent to modifying the average contrast.**
- **the contrast of the border layer δρ**

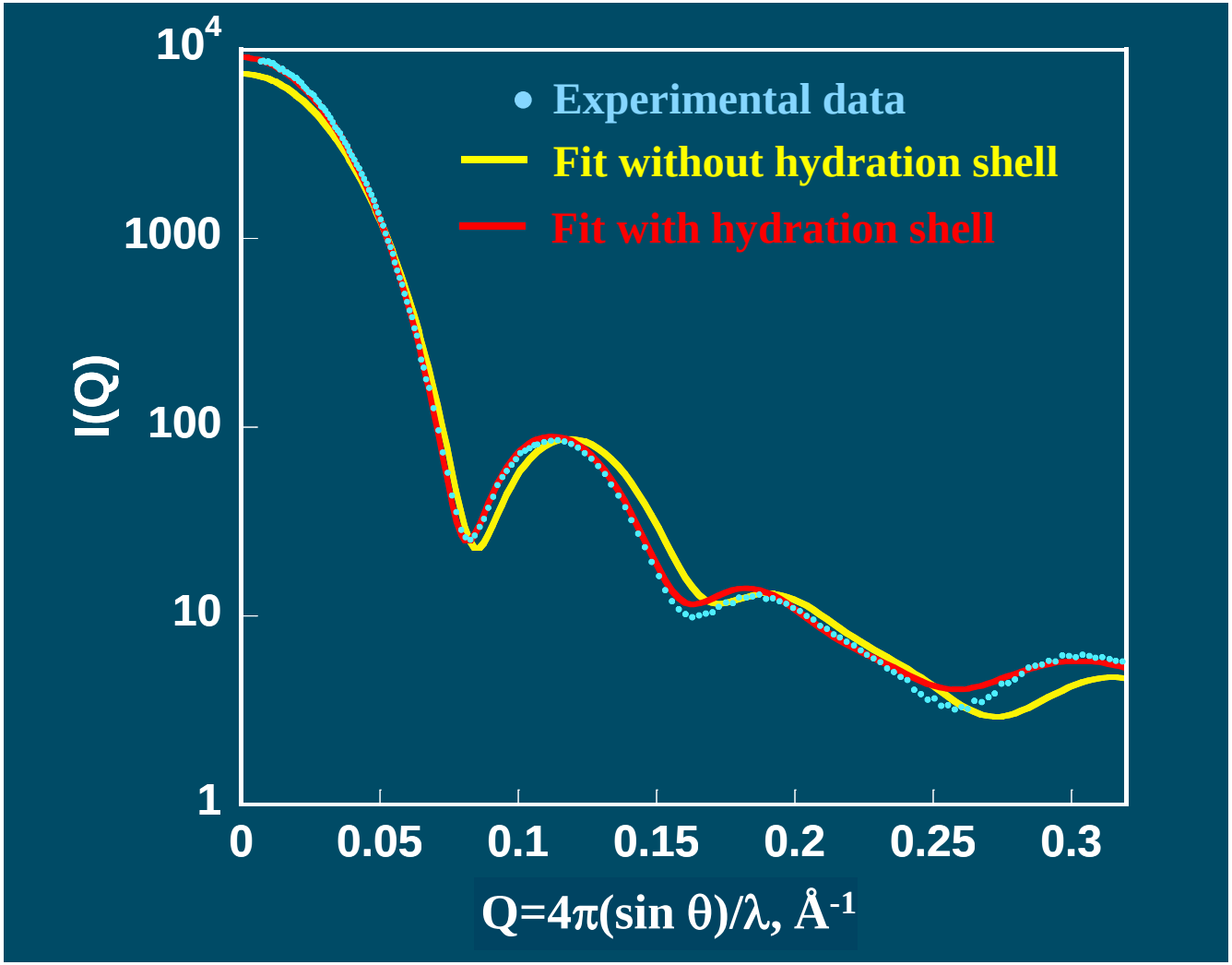
to minimize the discrepancy χ :

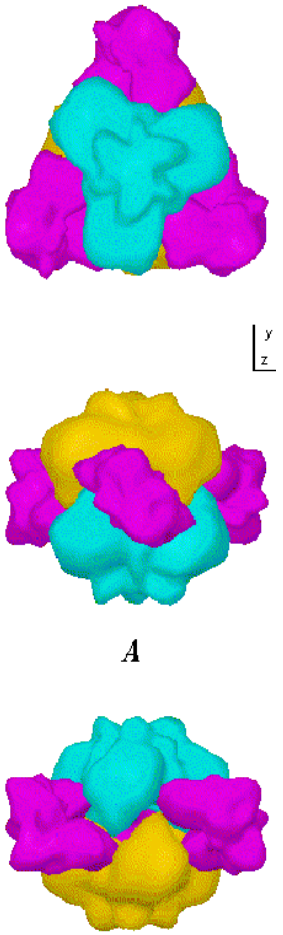
$$\chi^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left[\frac{I_{\exp}(Q_i) - scale * I_{calc}(Q_i)}{\sigma_{\exp}(Q_i)} \right]^2$$

REF: Svergun, Barberato & Koch (1995), Appl. Cryst., 28, 768 -773

Effect of the hydration shell

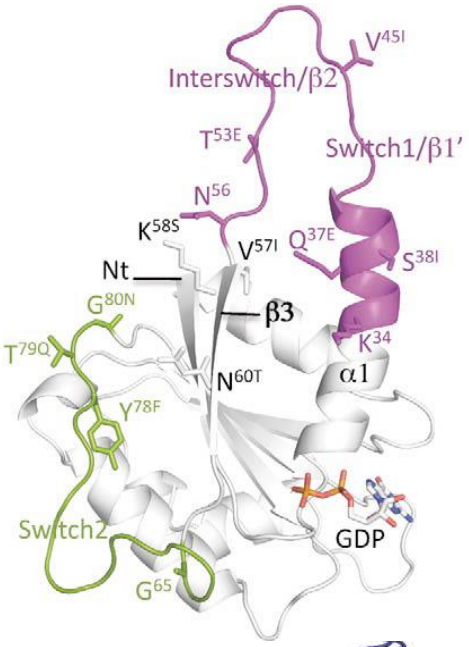
T state of *E. coli* allosteric ATCase

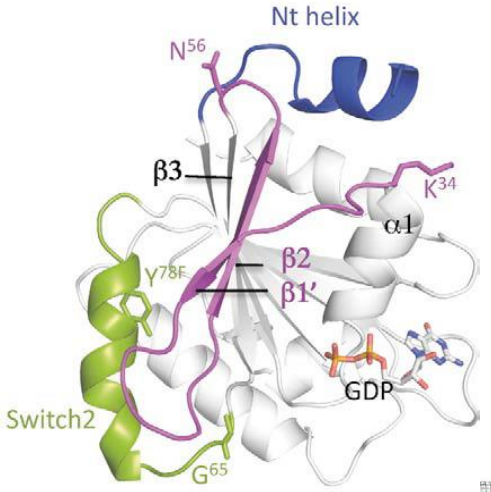




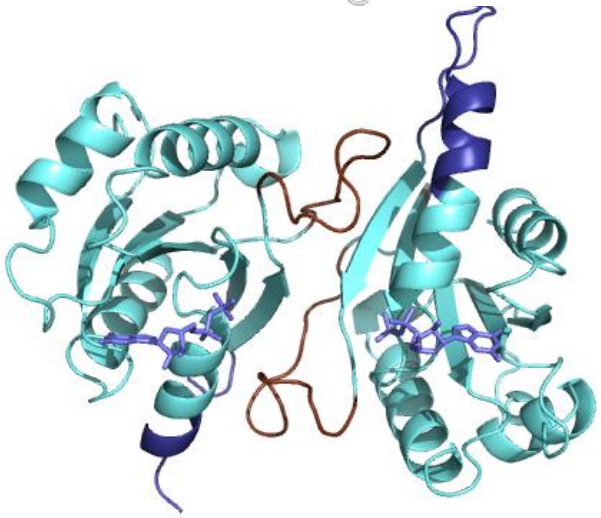
Crysol application : Arf6

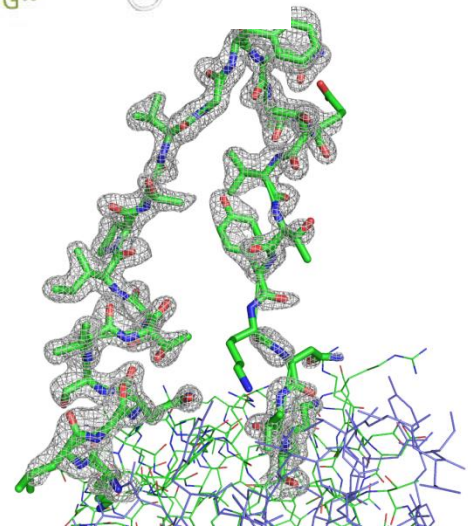
V. Biou et al., J.Mol.Biol (2010), 402(4), 696-707





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



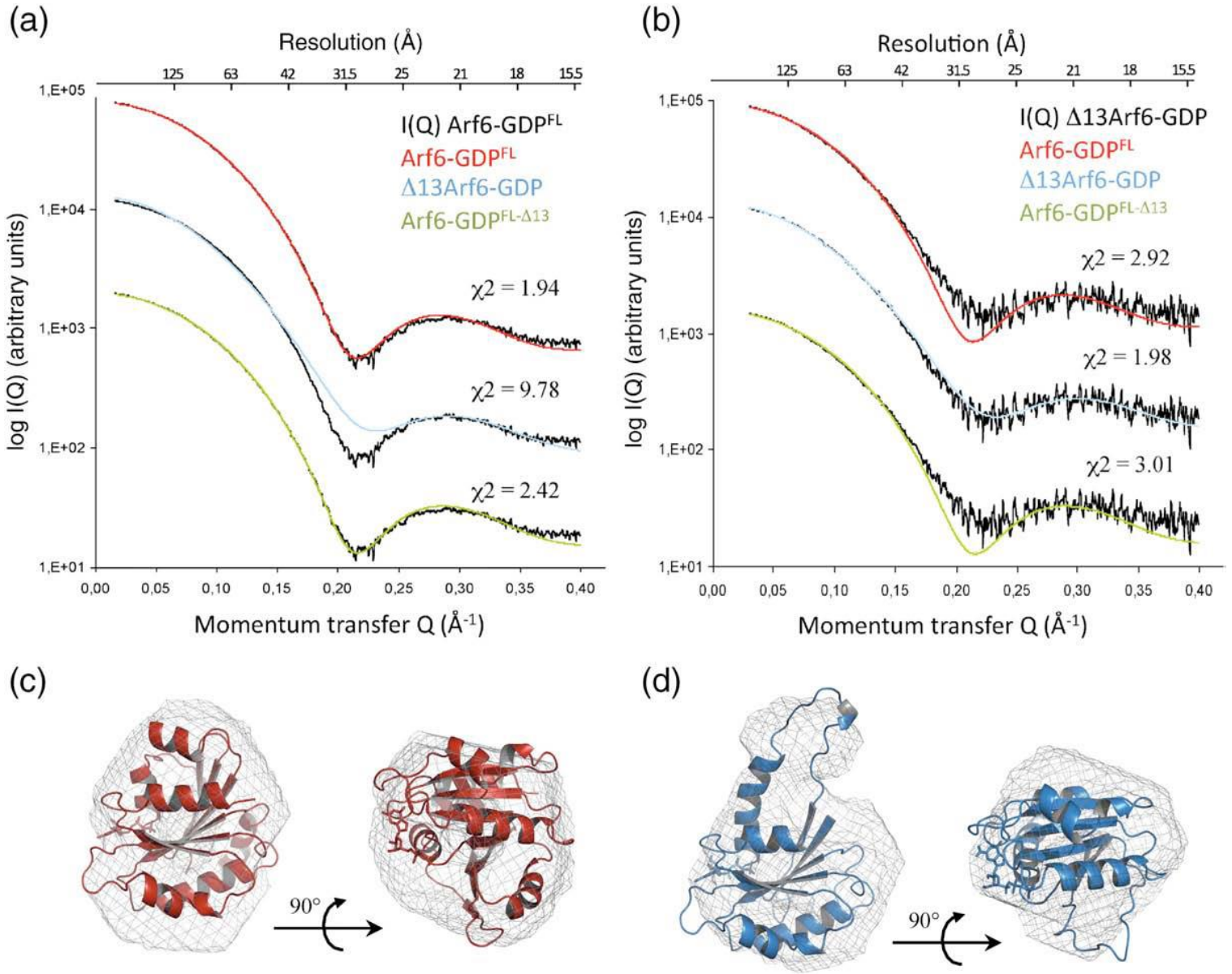


Δ13-Arf6 : a protruding loop in the crystal structure

Question : is the unfolded loop a crystal artifact ?

Crysol application : Arf6

V. Biou et al., J.Mol.Biol (2010), 402(4), 696-707



From an atomic structure to a solution scattering pattern

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

Grishaev A, Guo L, Irving T, Bax A. (2010) **AXES** Improved Fitting of Solution X-ray Scattering Data to Macromolecular Structures and Structural Ensembles by Explicit Water Modeling. J. Am. Chem. Soc. 132**,** 15484-6.
Use explicit water modeling solvation layer, robust fitting approach
http://spin.niddk.nih.gov/bax/nmrserver/saxs1/

J. Bardhan, S. Park and L. Makowski (2009) **SoftWAXS**: a computational tool for modeling wide-angle X-ray solution scattering from biomolecules J. Appl. Cryst. 42, 932-943
A program to compute WAXS,
Upon request

Schneidman-Duhovny D, Hammel M, Sali A. (2010) **FoXS**: a web server for rapid computation and fitting of SAXS profiles. Nucleic Acids Res. 3**8** Suppl:W540-4.
Debye-like computation, web server based. Hydration taken into account by "inflating" the volume of surface atoms.
http://modbase.compbio.ucsf.edu/foxs/

Knight C. J. and S. Hub J. S. (2015) **WAXSiS**: a web server for the calculation of SAXS/WAXS curves based on explicit-solvent molecular dynamics.
Nucleic Acids Res. 43 Suppl: W225-30.
http://waxsis.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

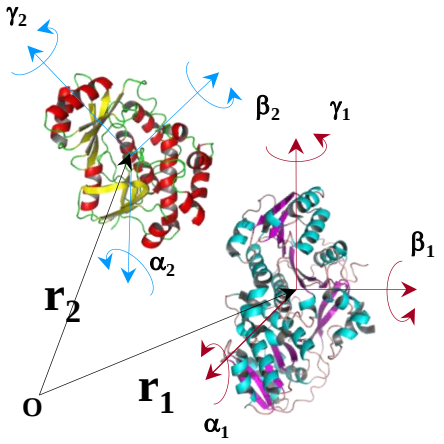
The objective is to find the relative orientation and position of each subunit that gives a good agreement with the SAXS data of the complex.

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

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

$$A^{(k)}(\vec{q}) = \exp(i.\vec{q}.\vec{r}_k) \prod\prod (\alpha_k.\beta_k.\gamma_k)[C^{(k)}(\vec{q})]$$

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.

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

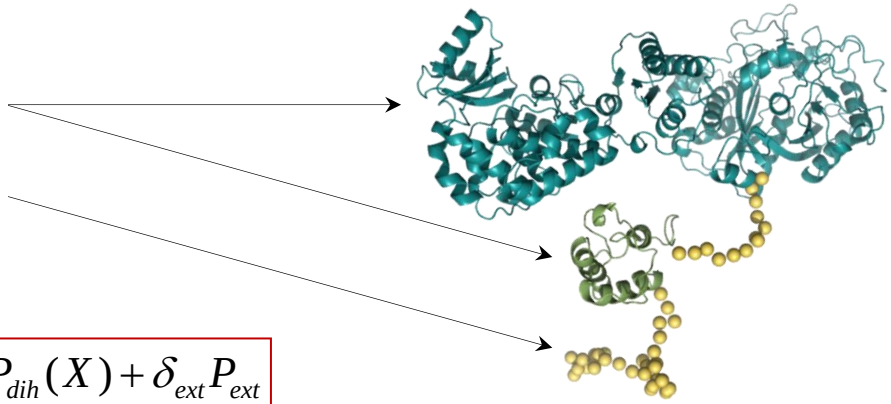
Petoukhov & Svergun (2005). Biophys. J., 89, 1237-1250.

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*



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

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

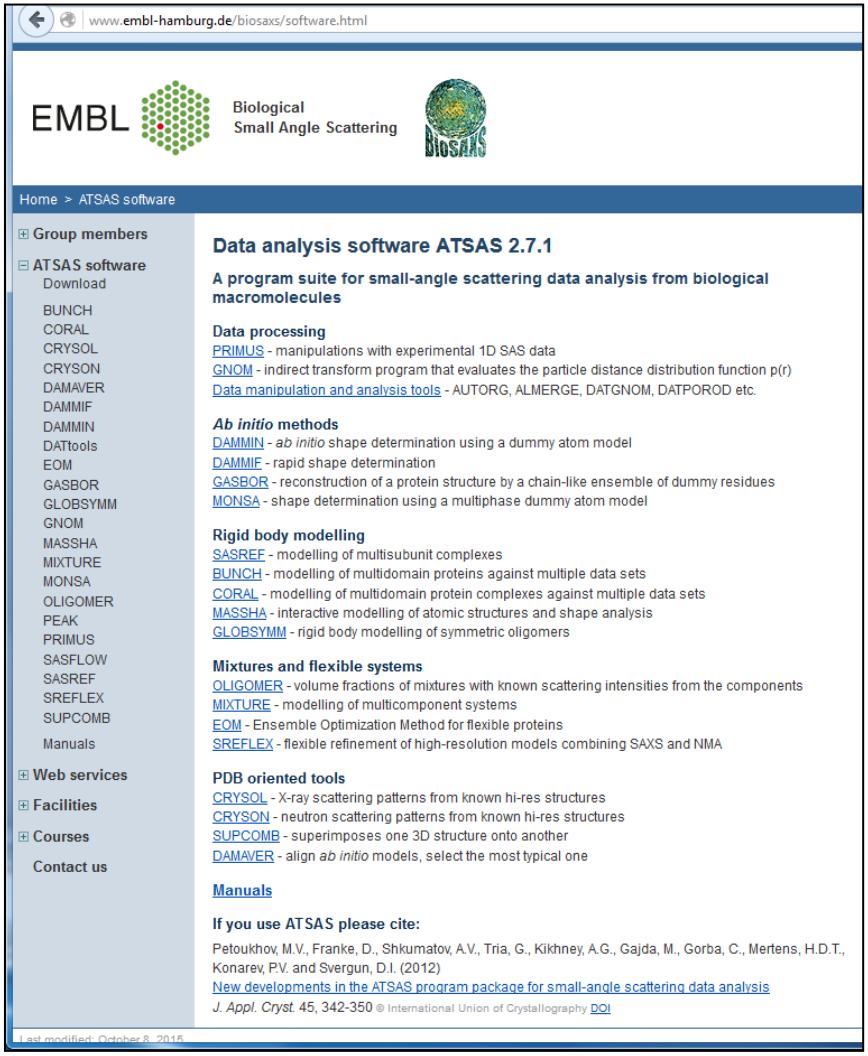
The algorithm of minimization is the same used with SASREF with a penalty function including the steric clashes Pcross, the dihedral angle Pang and Pdih, and the compactness of the loop Pext. The possibility to give information about contacting residues from other experiences is also added.

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

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

ATSAS package and ATSAS online

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



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

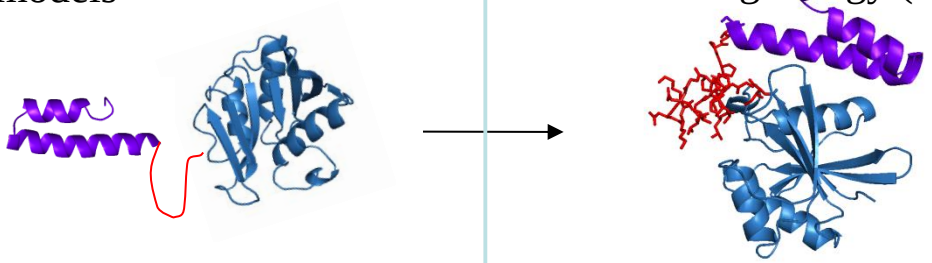
DADIMODO : rigid body refinement *vs.* SAXS / NMR data

Collab : Christina Sizun & François Bontems (ICSN, Gif sur Yvette))
Evrard et al. (2011), *J. Appl. Cryst.*, **44**:1264-1271.

Modelling approach : complete atomic model

Full structure initiated with :

- Crystal or NMR domain structures
- Homology models

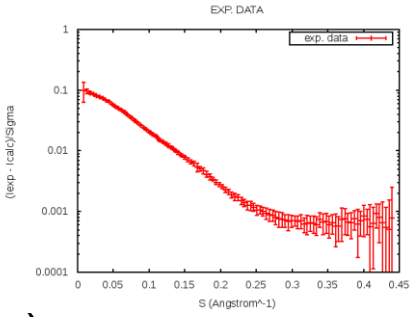


External information:

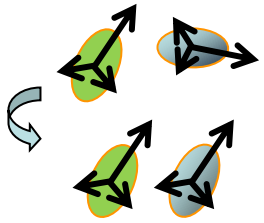
- Sequence
- Sub-parts moved as rigid-bodies (user-defined)
- A correct stereochemistry is maintained at all steps by minimizing energy (Amber 99 Force Field)

Experimental data:

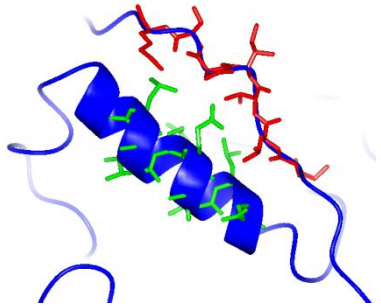
- SAXS
- NMR
  RDC
  ADR (chem. shift map.)



SAXS score



RDC score

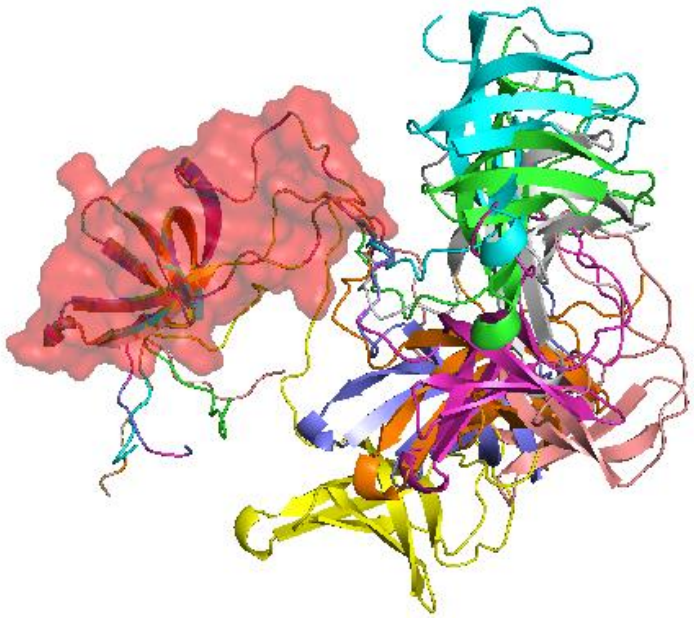


ADR score

Optimisation of the structure via a genetic algorithm

Dadimodo example : F45 from S1 protein

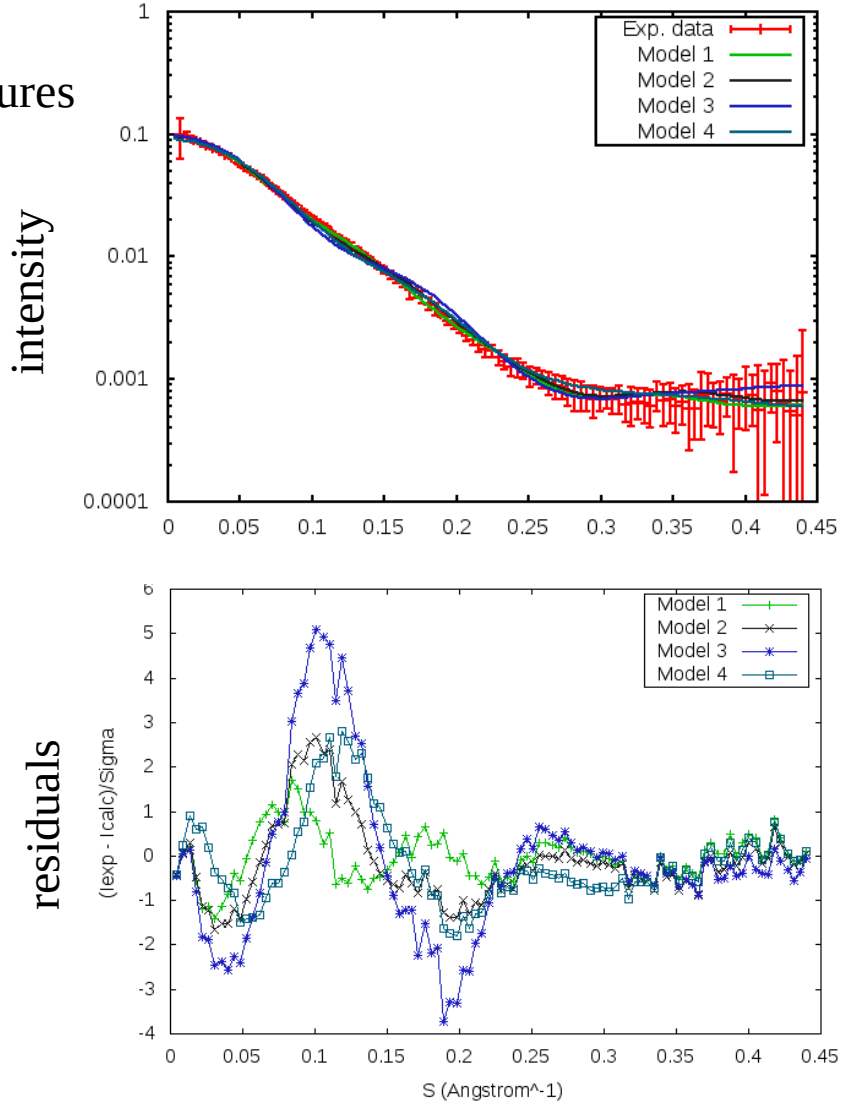
Structure:
- one polypeptide chain
- two rigid domains (D4 & D5) with known structures
- 1 linker
- 2 flexible parts (N-term and C-term)



Starting models population

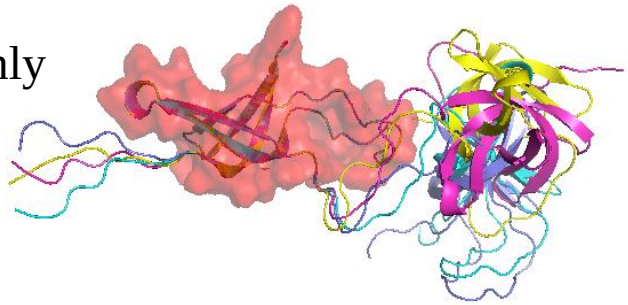
- **obtained after running 50 steps Dadimodo with no selection pressure**

Initial misfit to the SAXS data:

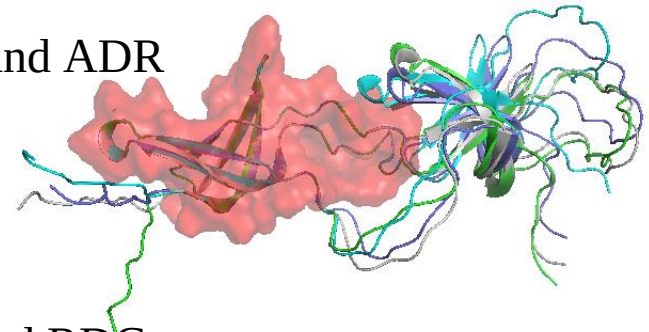


Dadimodo example : F45 from S1 protein

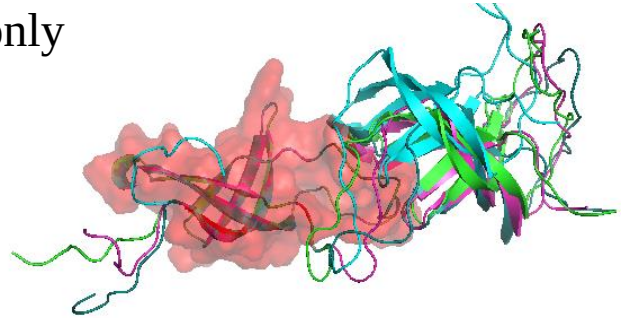
vs. SAXS only



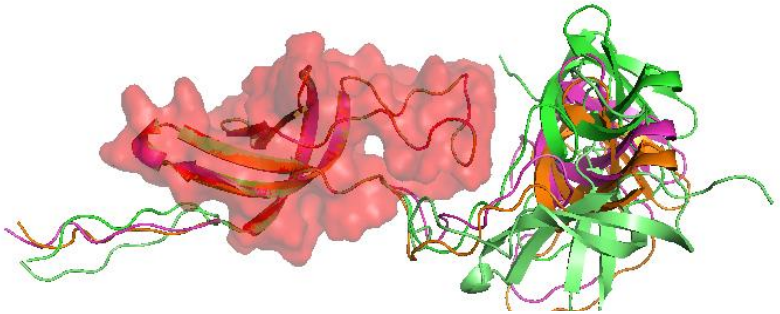
vs. SAXS and ADR



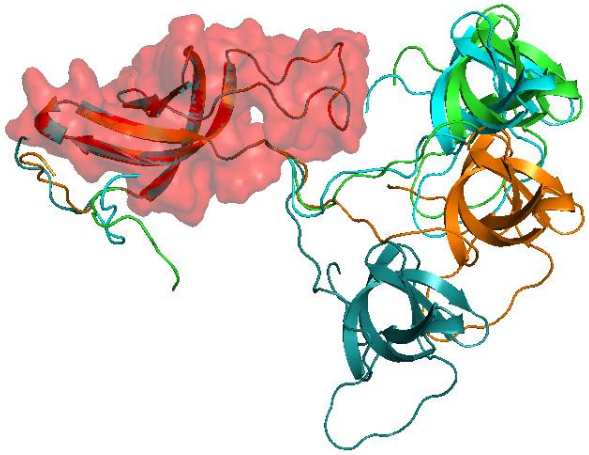
vs. ADR only



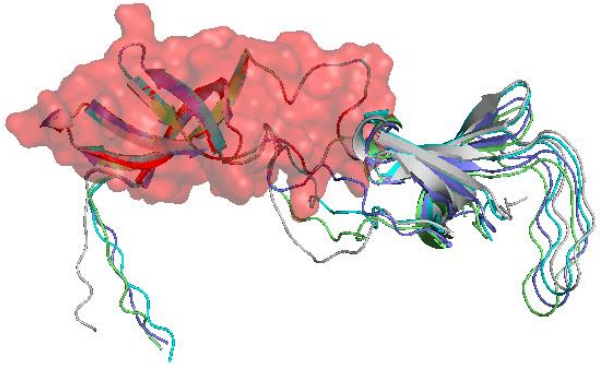
vs. SAXS and RDC



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, a*b 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 <u>independently</u>.
- ✓ Otherwise :

IN

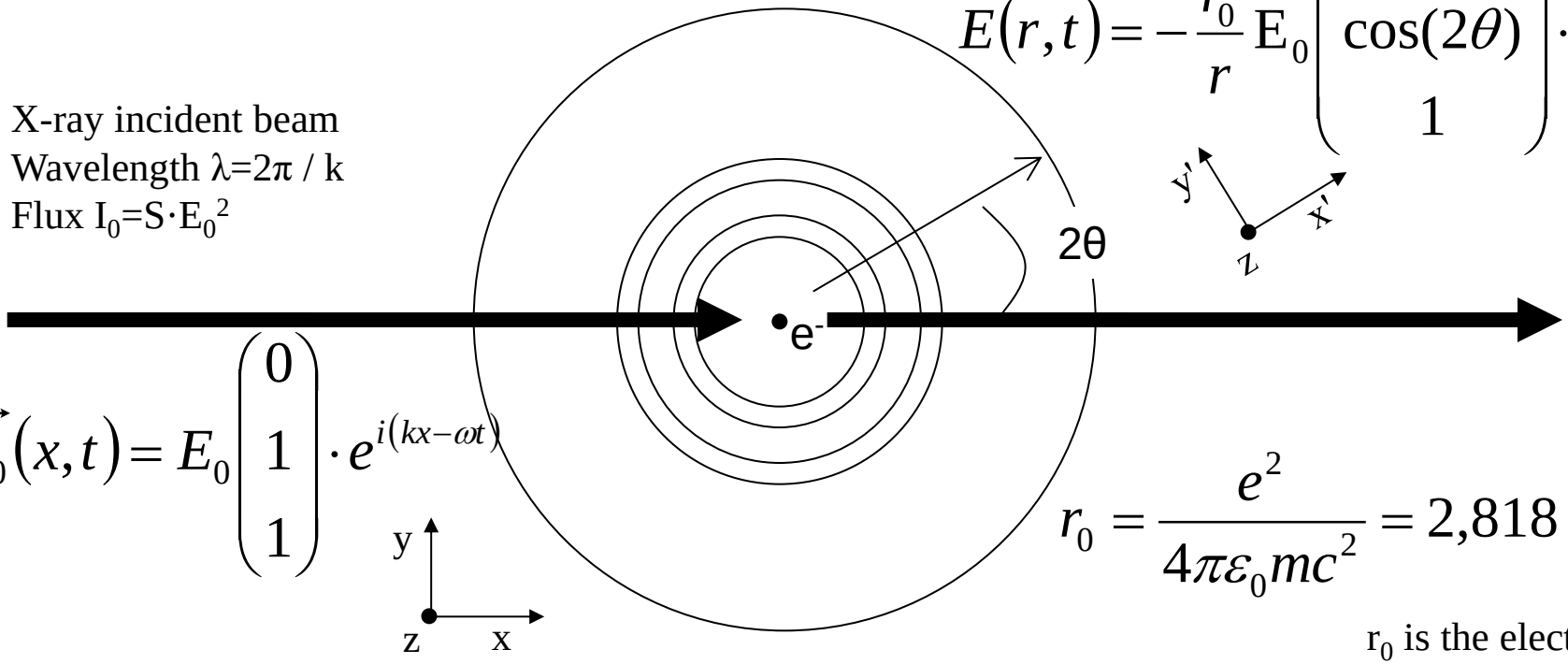
OUT

Elastic Thompson scattering by an electron

- **What scatter X-rays are <u>the electrons</u>**

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

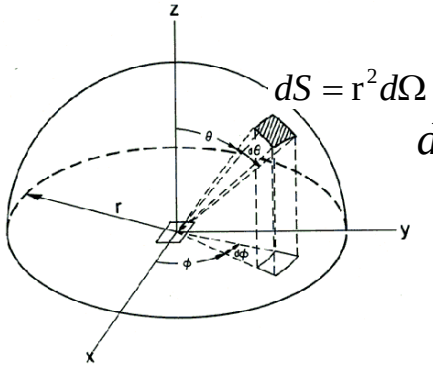
X-ray incident beam
Wavelength $\lambda = 2\pi / k$
Flux $I_0 = S \cdot E_0^2$



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

$$r_0 = \frac{e^2}{4\pi\varepsilon_0 mc^2} = 2{,}818\ 10^{-15}\, m$$

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

Data analysis

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

Molecular modeling

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

Nothing is known
- low resolution shape

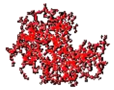
Structures of subunits available
- molecular modeling rigid body against SAXS data

Structures with missing loop or flexible parts
- molecular modeling of missing parts against SAXS data

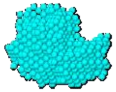
From an atomic structure to a solution scattering pattern

$$I_{th}(q) = \left\langle \left| A_a(\vec{q}) - \rho_s A_s(\vec{q}) + \delta\rho_b A_b(\vec{q}) \right|^2 \right\rangle_\Omega$$

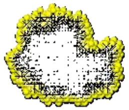
$A_a(\mathbf{q})$ = molecular scattering amplitude in vacuum



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



$A_b(\mathbf{q})$ = 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 :

$$I_{calc}(q) = \sum_{l=0}^{L} \sum_{m=-1}^{l} \left| A_{lm}(q) - \rho_0 C_{lm}(q) + \delta\rho B_{lm}(q) \right|^2$$

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

- the general scale of $I_{calc}(q)$
- the total excluded volume V, which is equivalent to modifying the average contrast ρ_0
- the contrast of the border layer $\delta\rho$

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

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

3D shape reconstructions from SAXS data with DAMMIN

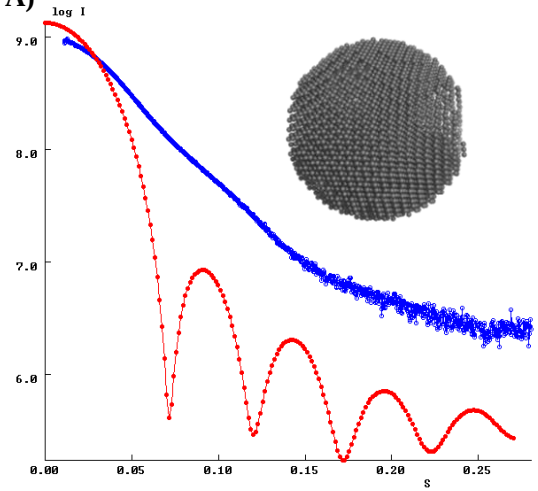
Ab initio shape modelling : nothing is known excepted the curve !

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 = $D_{max}/2$ with N scattered beads ($r_0 << R$)
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)

A)

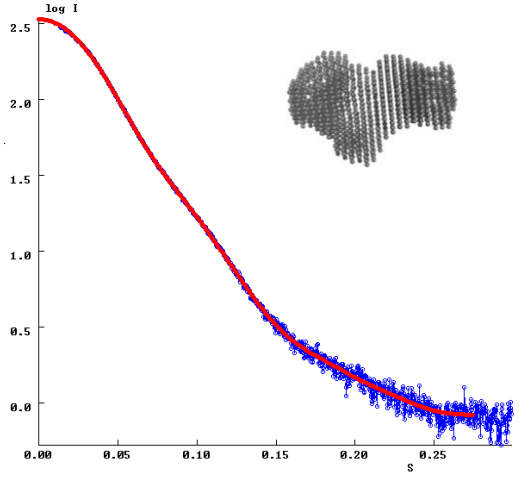


$$f(X) = \chi^2\left[I(q)_{\exp}, I(q,X)\right] + \alpha P(X)$$

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.