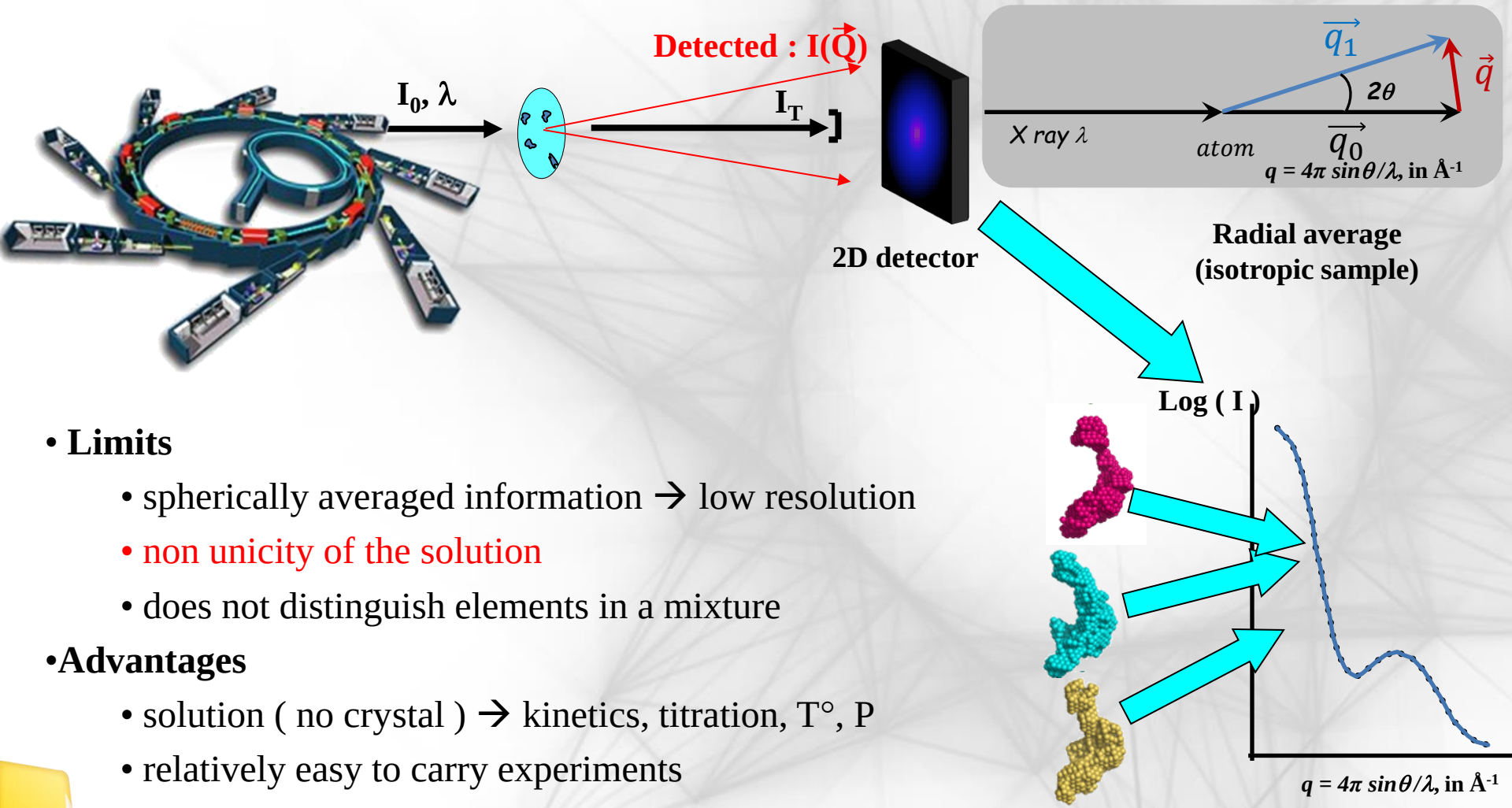
Solution X-ray Scattering from Biological Macromolecules

Aurélien Thureau
Beamline SWING, Synchrotron SOLEIL, Saint-Aubin, France

INTRODUCTION

Principles of Small Angle X-ray Scattering 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 obtained from a scattering curve

- biophysical parameters (**size** and **shape type**)
- **molecular mass**, **oligomerization state** and **volume**.

Biophysical informations derived directly from the SAXS curve

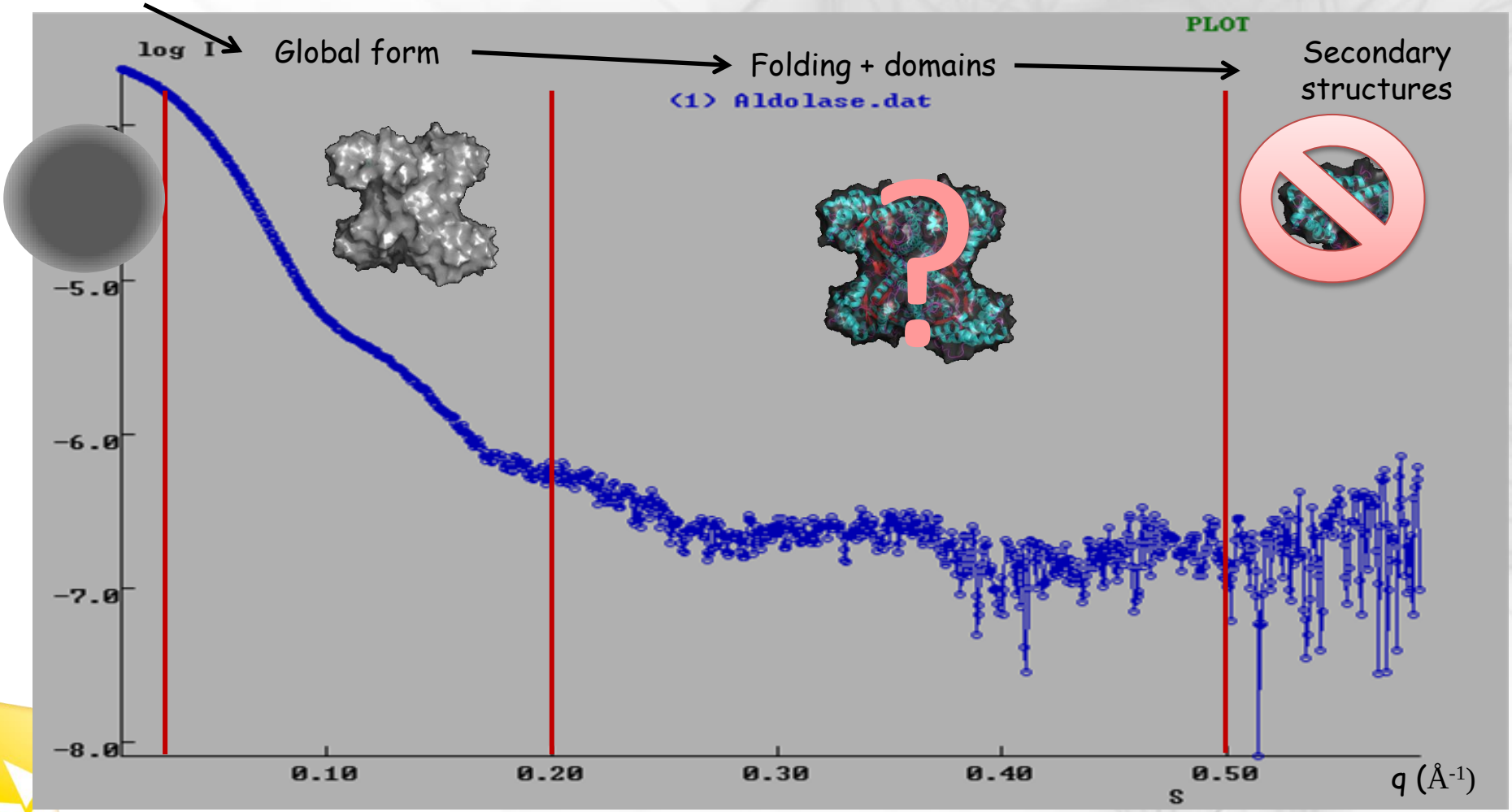
- possible **low resolution** molecular shape (ab initio methods)

3D structural modeling
→ compatible models with SAXS data

- direct comparison with high resolution model
- possible model of (un)structured missing parts
- rigid body of complex

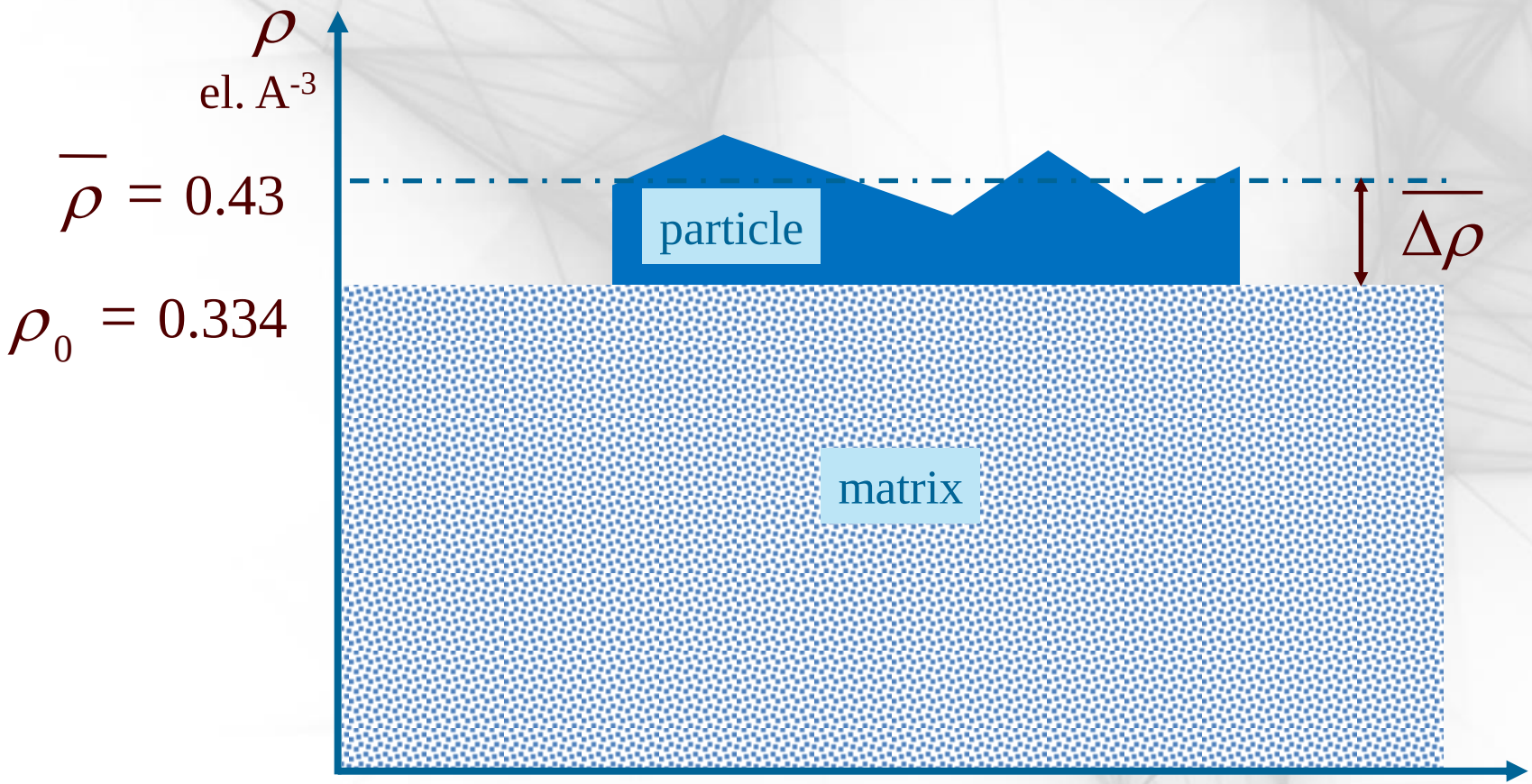
NOT a unique model,
NO electron density map.

What may solution scattering yield?

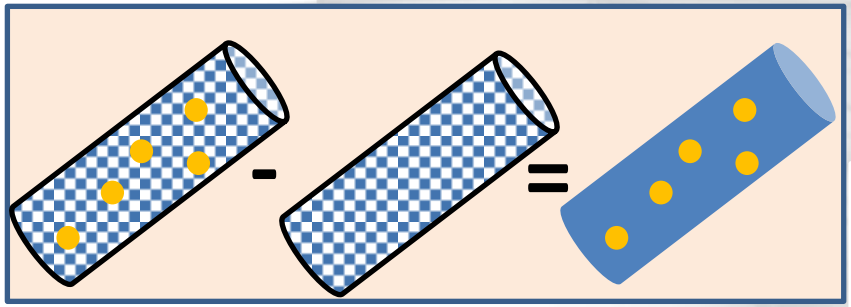


Particles in a matrix (or buffer)

- 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.

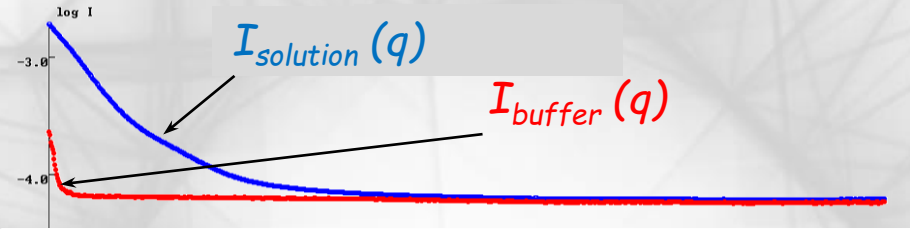


A SAXS curve results from a pair of measurements : solution & buffer



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

Log scale

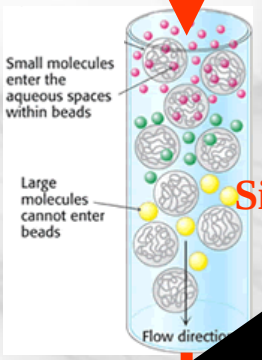


To obtain scattering solely from the contrasting particles, intrinsic solvent scattering must be measured **very accurately** and subtracted, which also permits to subtract contribution from parasitic background (slits, sample holder etc) which should be reduced to a minimum.

Solution Sampler / SEC-HPLC

Flow rate 300 µl/min

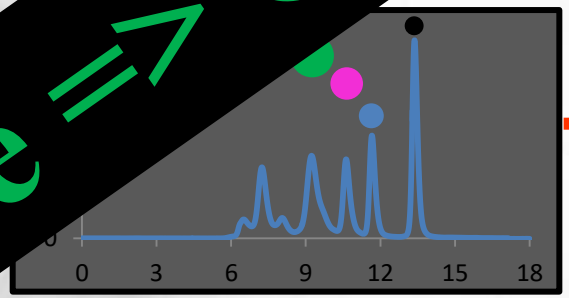
- Monodisperse solution
- Aggregation is eliminated
- Oligomeric conformation
- Equilibrium states c… …ated
- Perfect backgro…
- Automatic c…



Size Exclu…

Incident X-ray

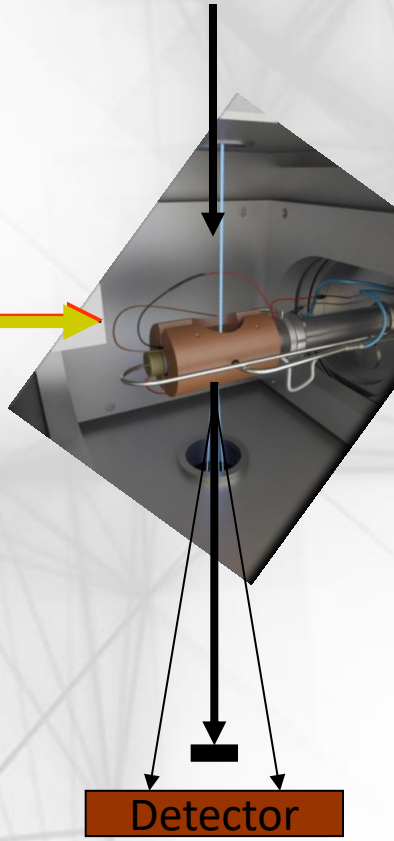
Samples injection:



…e 70 µl/min

Pure sample

- Small volumes (~ 10 µl)
- No dilution
- High rate (~2 minutes/sample)



Good Sample ⇒ Good Data

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).

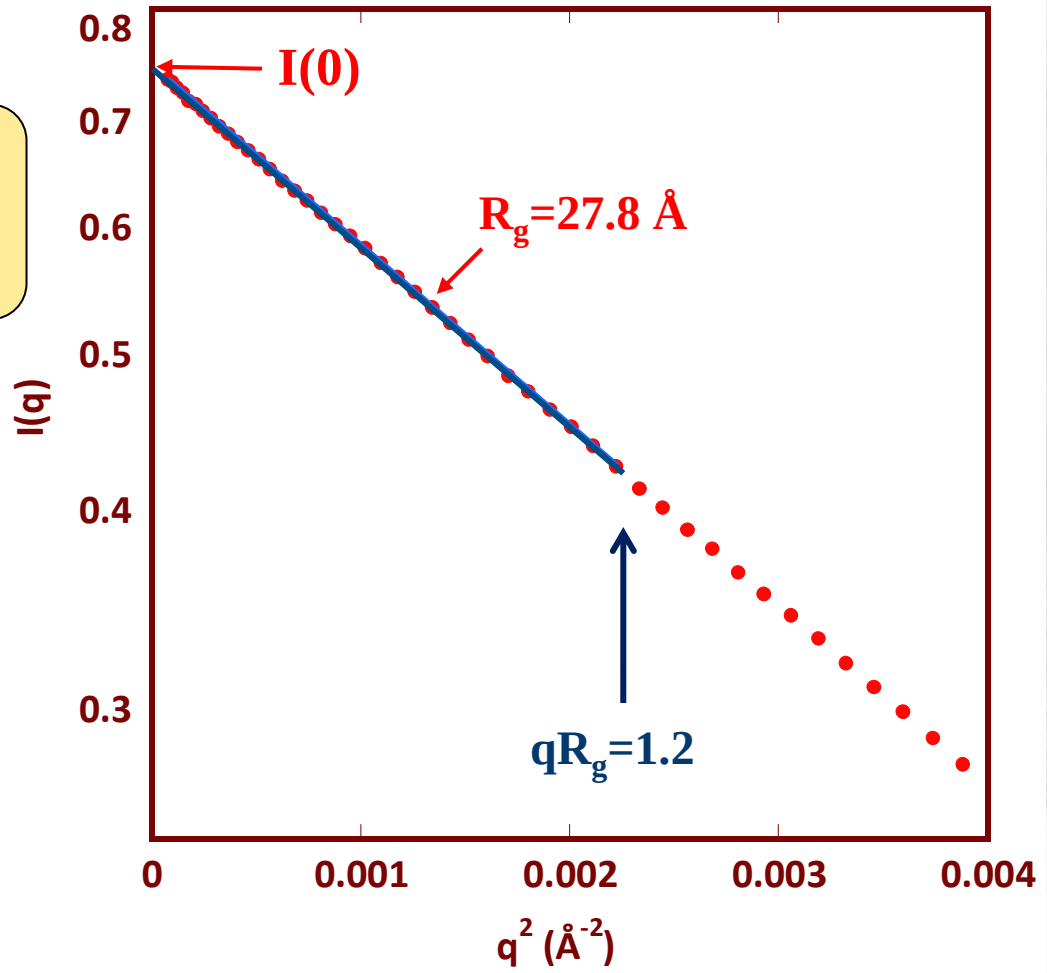
Data Analysis : Guinier law

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 :

$0 < qR_g < 1$ for a solid sphere

$0 < qR_g < 1.3$ rule of thumb for a globular protein

ideal monodispersed

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

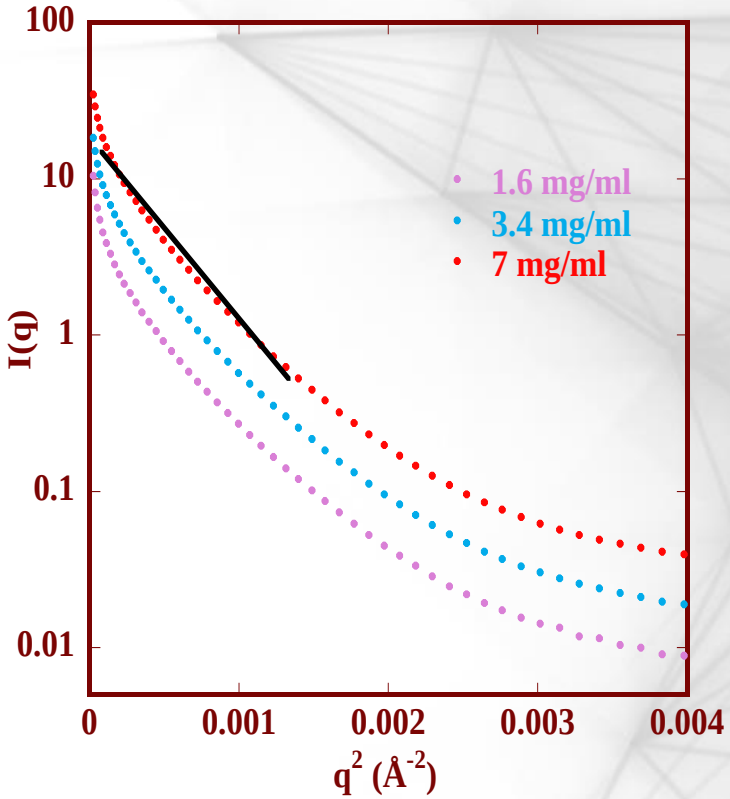
For globular proteins : $R_g\,(Å) \approx 6.5 * M^{\frac{1}{3}}$, M in kDa
For unfolded proteins : $R_g\,(Å) \approx 8.05 * M^{0.522}$

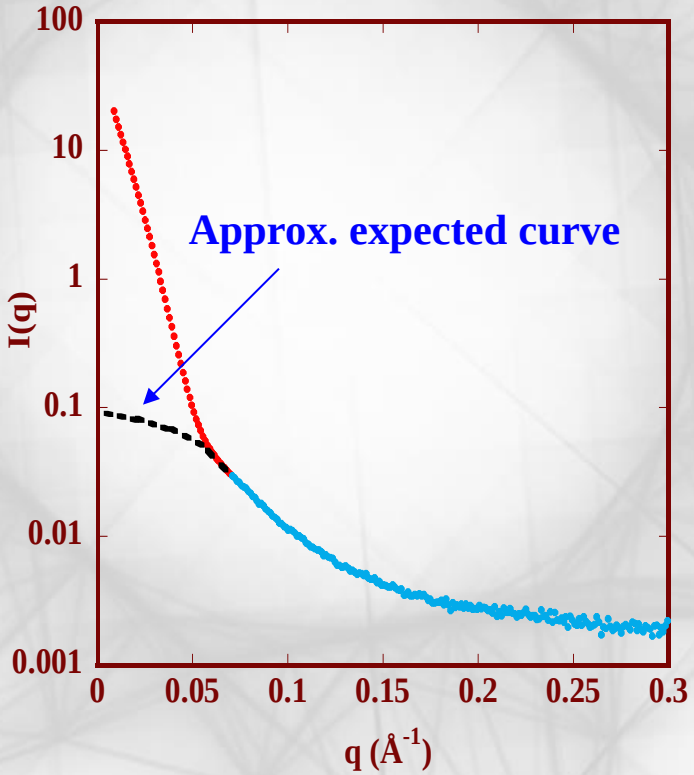
Bernado et al. (2009), Biophys. J., 97 (10), 2839-2845.

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

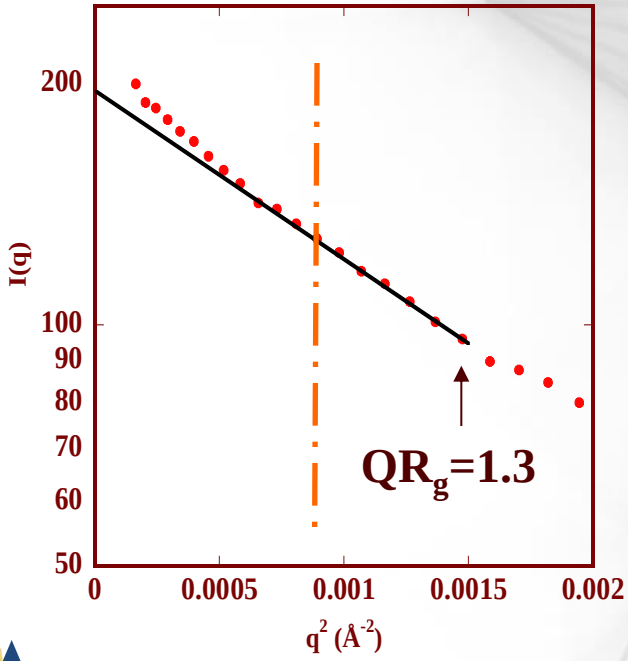
(Courtesy D. Durand, IBBMC, Orsay)

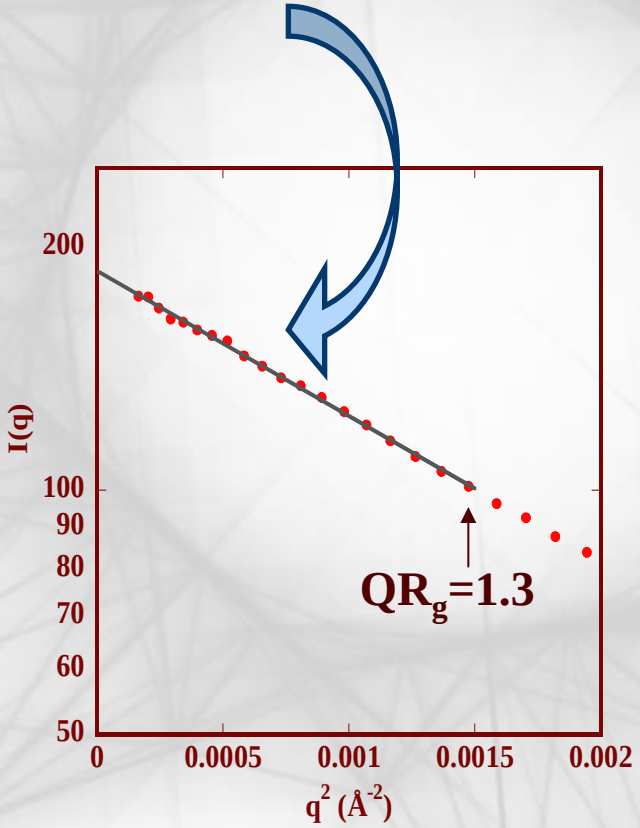
Evaluation of the solution properties

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

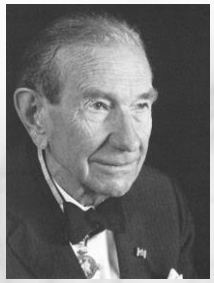
Guinier plot

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

Data Analysis

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

Kratky Plot

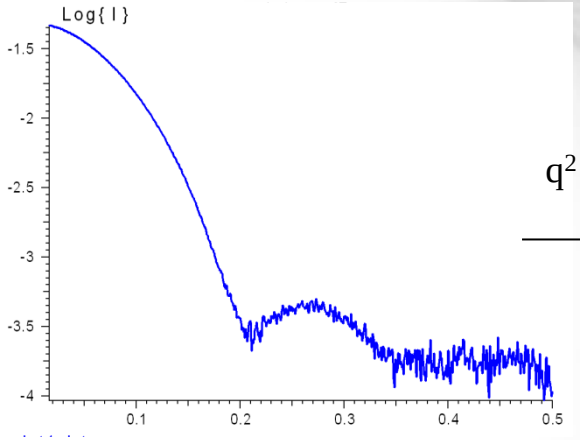


Prof. Otto Kratky
1902-1995
Graz, Austria

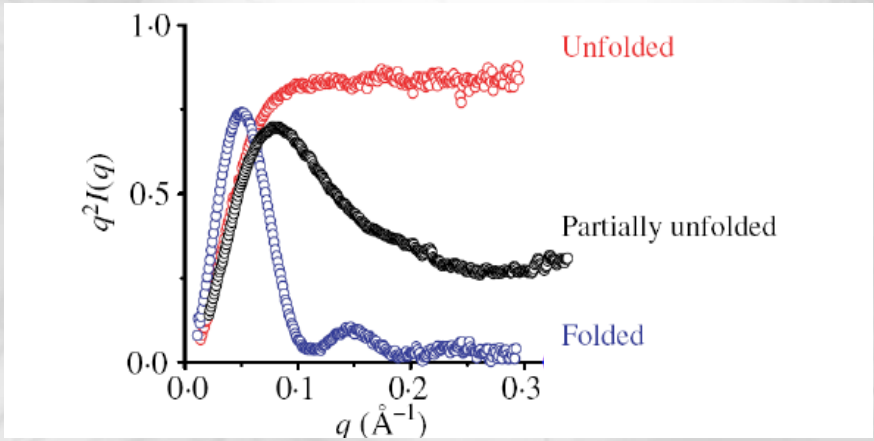
SAXS provides a sensitive means to ***evaluate the degree of compactness*** of a protein:

- 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



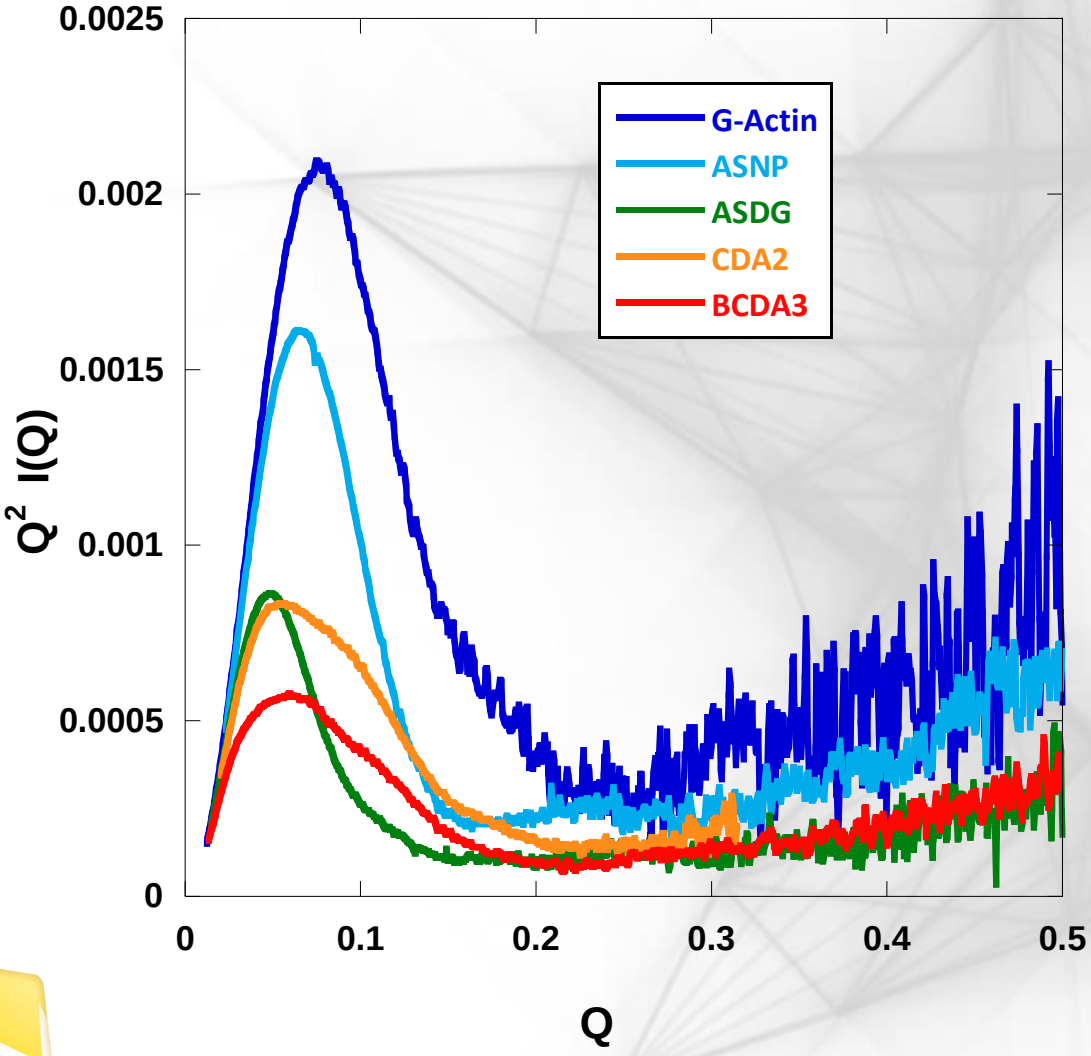
Putnam, D., et al. (2007) Quart. Rev. Biophys. 40, 191-285.

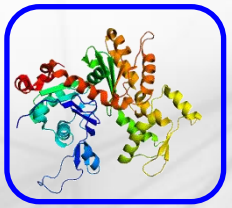
Folded particle : ***bell-shaped curve*** (asymptotic behaviour $I(Q) \sim Q^{-4}$)

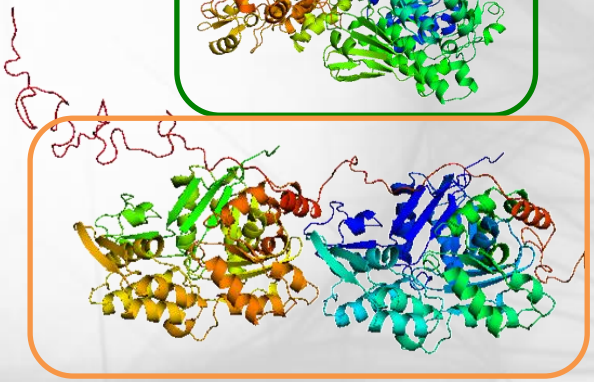
Random polymer chain : ***plateau*** at large q-values (asymptotic behaviour in $I(Q) \sim Q^{-2}$)

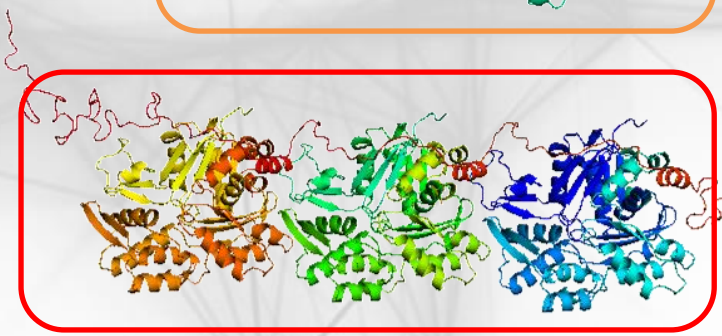
Extended polymer chain : ***increase*** at large q-values (asymptotic behaviour in $I(Q) \sim Q^{-1.x}$)

Kratky Plots of folded proteins









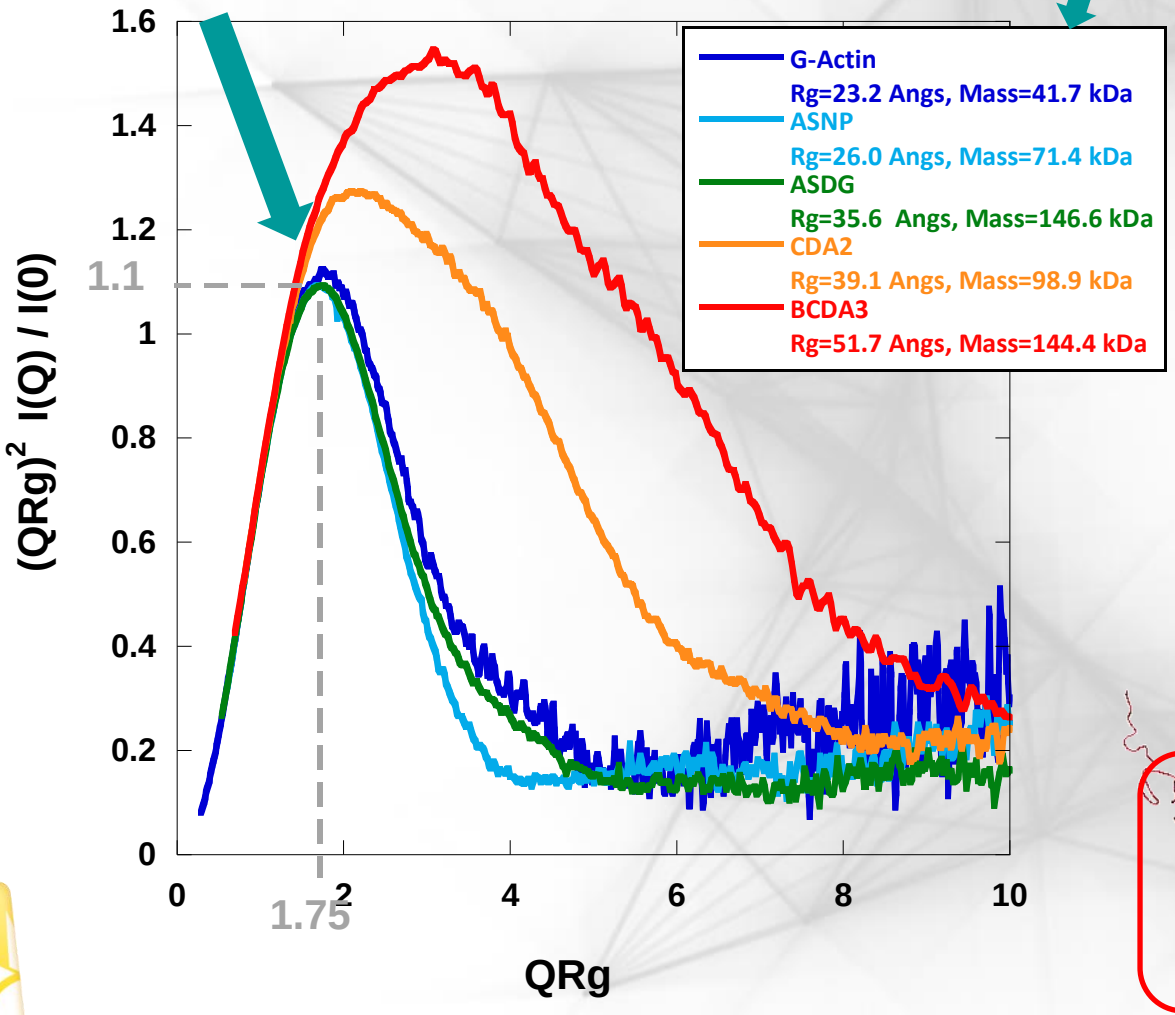
Folded proteins display a bell shape. Can we go further?

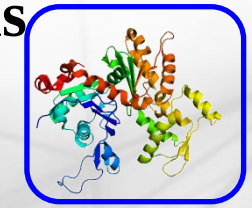
Dimensionless Kratky Plots of folded proteins

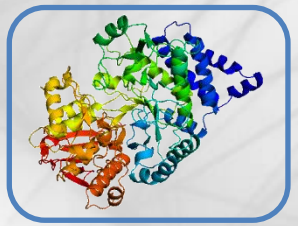
Introduced for biology in Durand et al. (2010), J. Struct. Biol. **169**, 45-53.

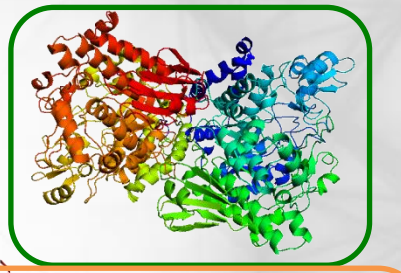
For globular structures, DLKPs fold into the same maximum

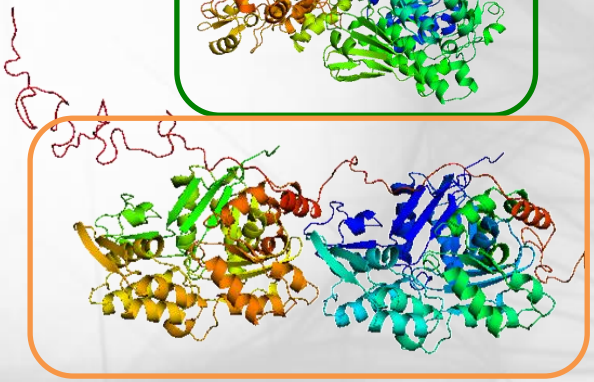
The relation $M_{Rg}(kDa) \approx (Rg / 6.5)^3$ only works for the globular structures, not the elongated

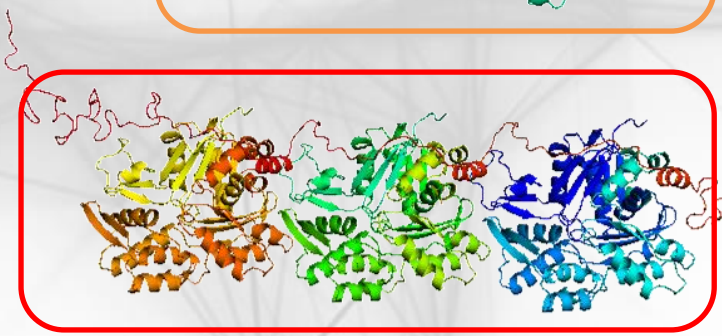








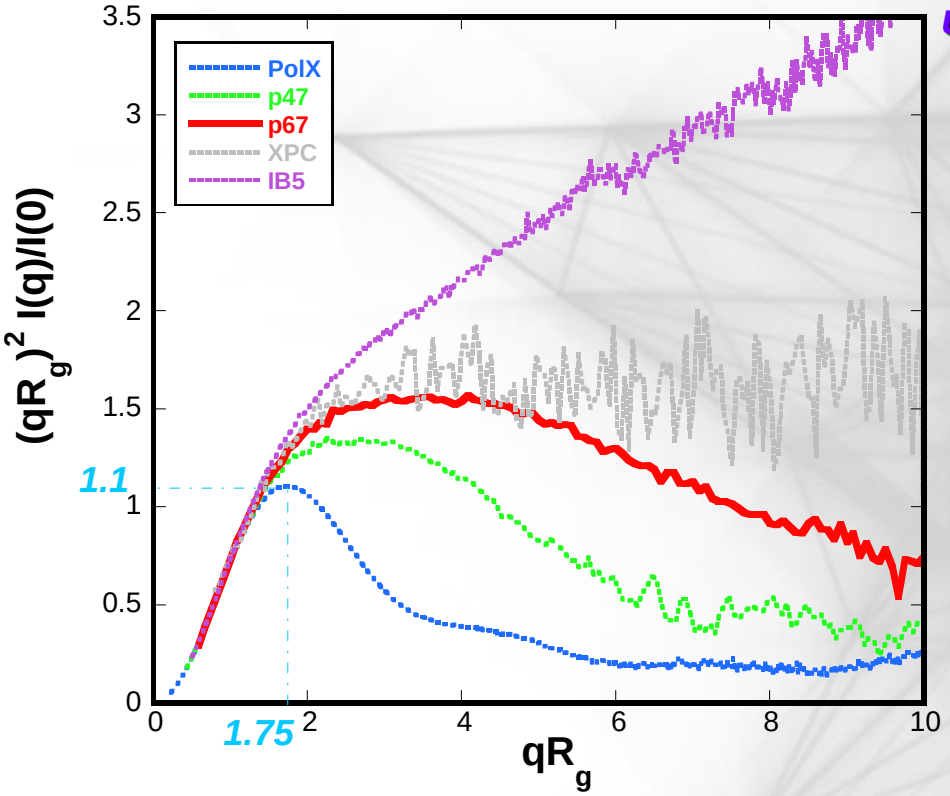




The maximum value on the dimensionless bell shape tells if the protein is globular.

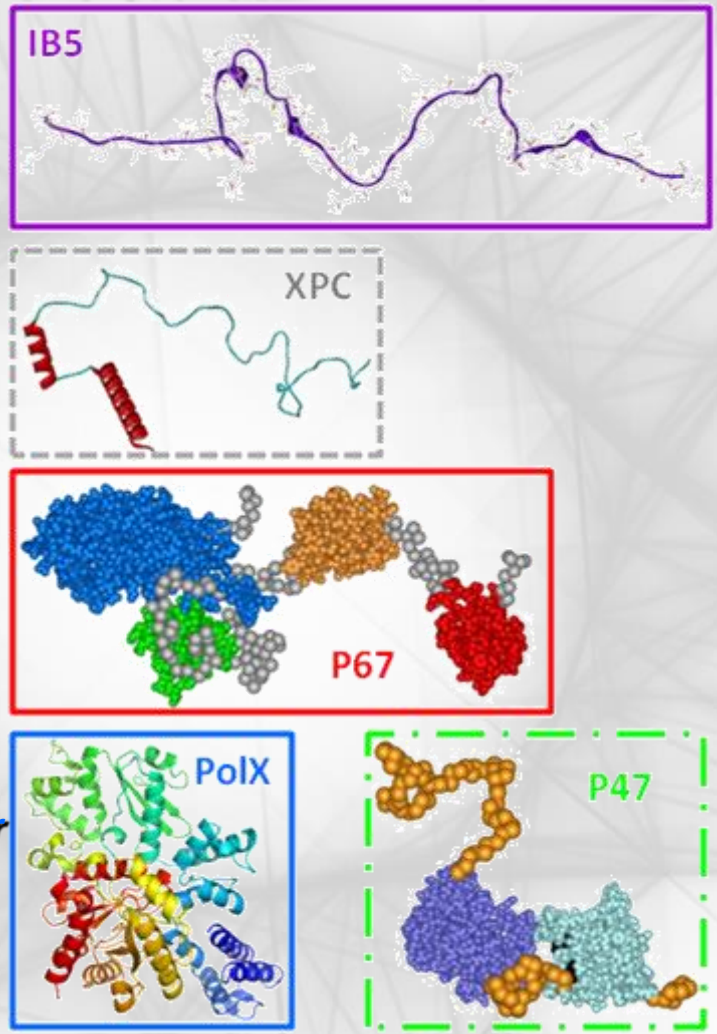
Dimensionless Kratky Plots of (partially) unfolded proteins

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.

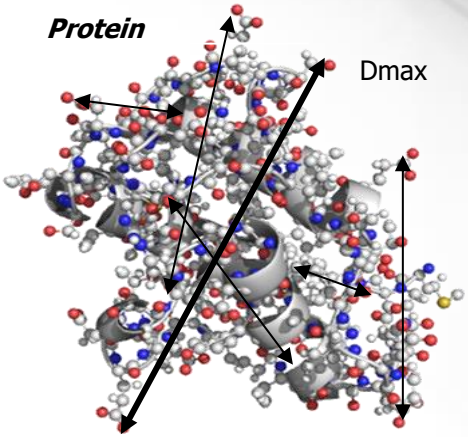
The curve increases at large Q as the structure extends.

Data Analysis

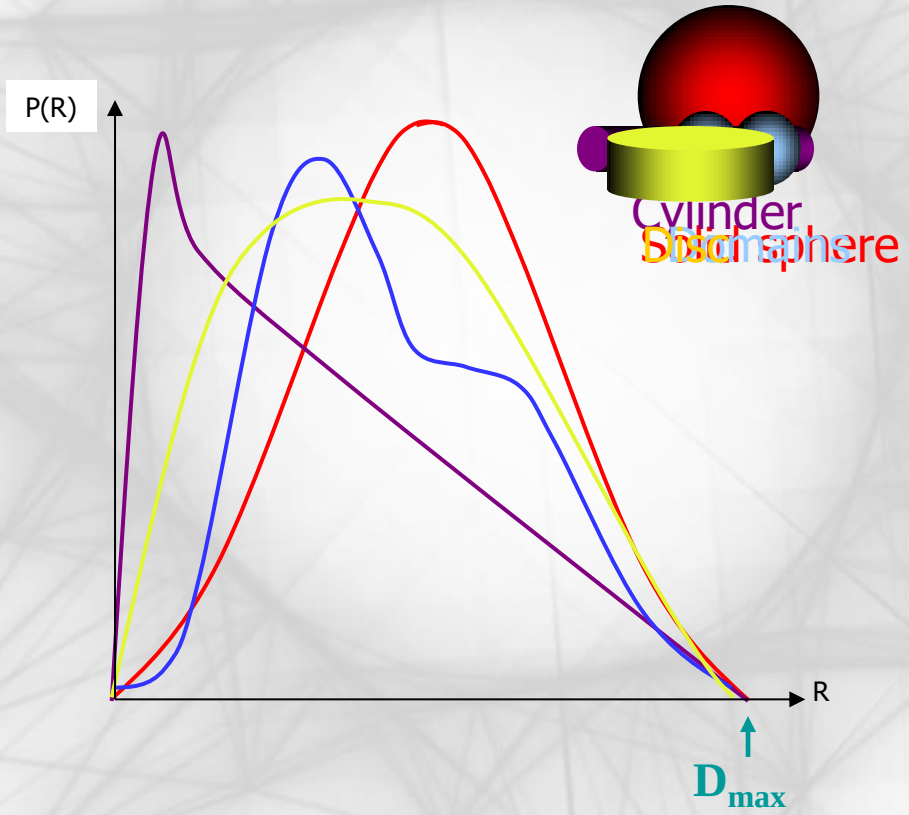
- Guinier Analysis
- Kratky plot : why is it so interesting ?
- « Real-space SAXS » : Distance correlation function P(r)

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 :

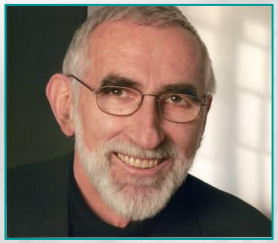
$$p(r) = \frac{r^2}{2\pi^2 \varphi r_e^2} \int_0^\infty q^2 I(q) \frac{\sin(qr)}{qr} dq$$

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

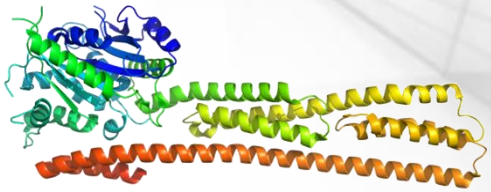
+ "Perceptual criteria" : smoothness, stability, absence of systematic deviations

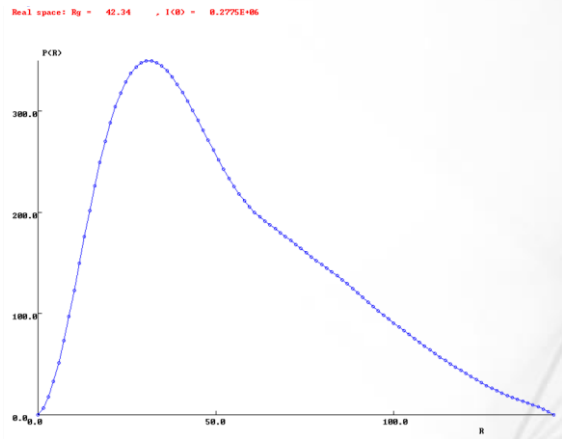
- Each criterium has a predefined weight
- The solution is given a score calculated by comparison with « ideal values »

Distance Distribution Function

Experimental examples

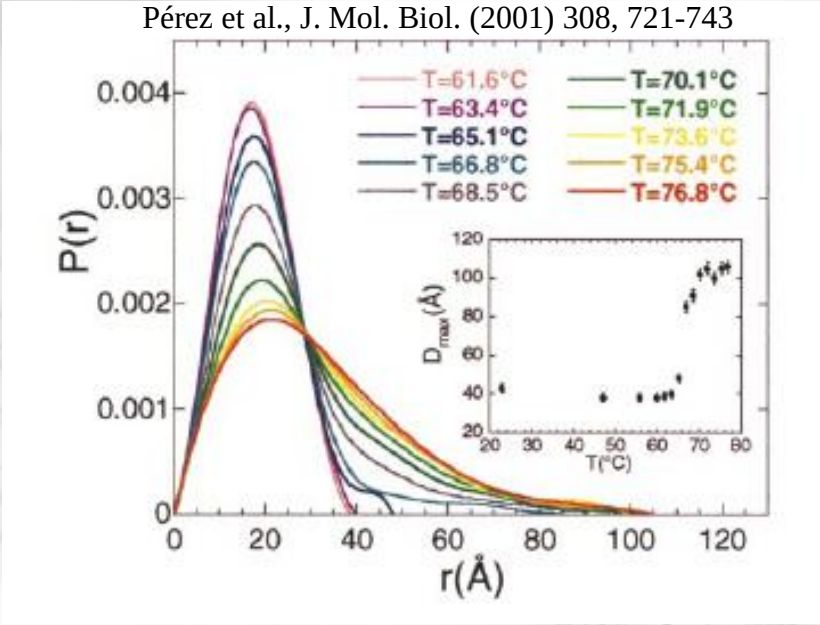
GBP1





Heat denaturation of Neocarzinostatin

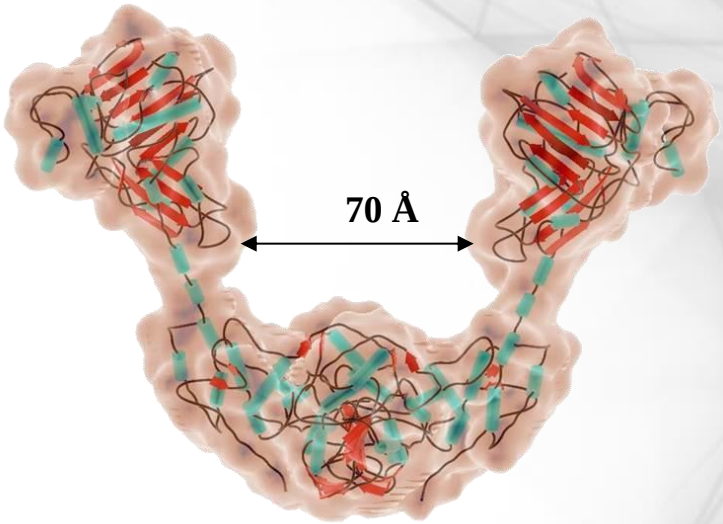
Pérez et al., J. Mol. Biol. (2001) 308, 721-743



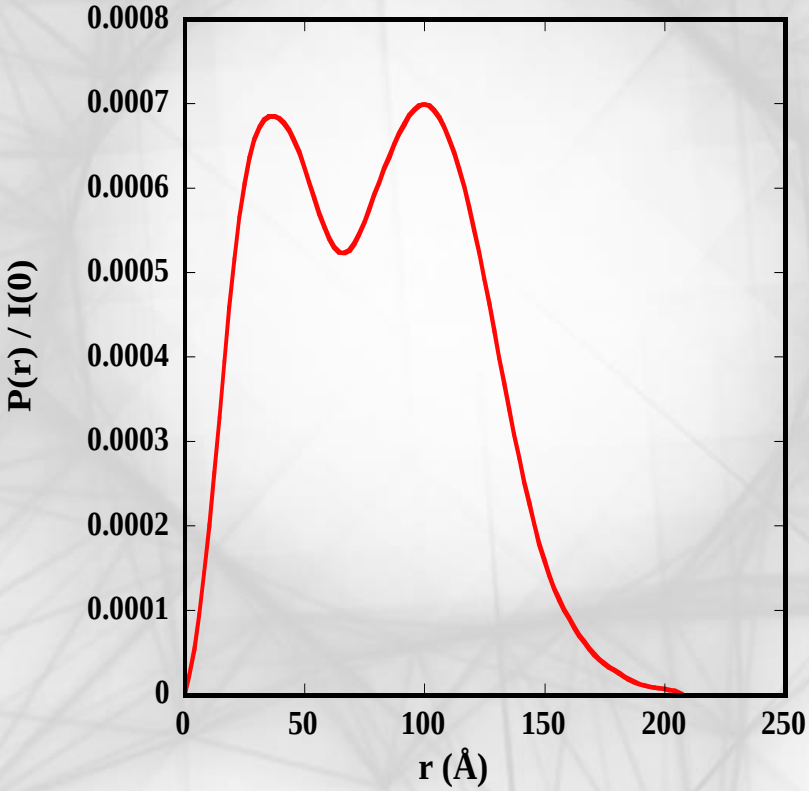
Distance Distribution Function

Experimental examples

Topoisomerase VI



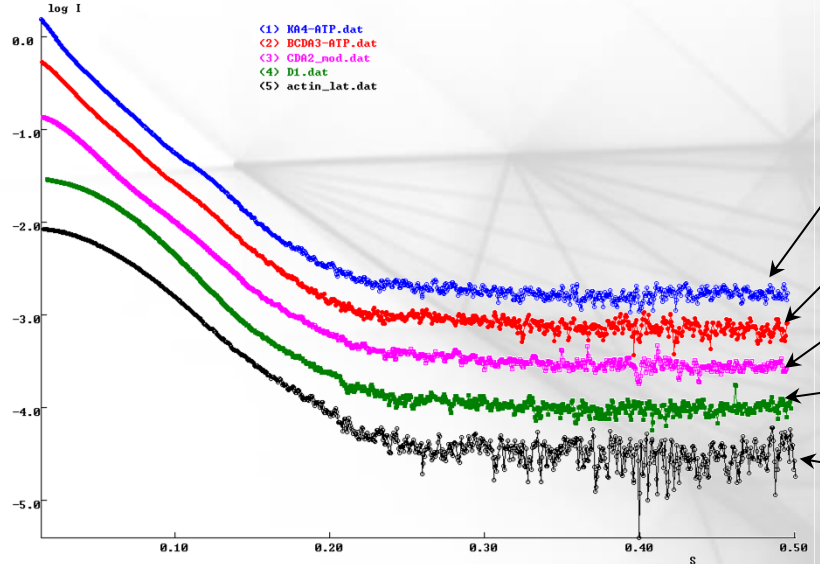
Bimodal distribution

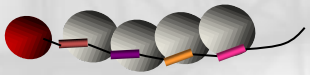


M. Graille et al., Structure (2008), *16*, 360-370.

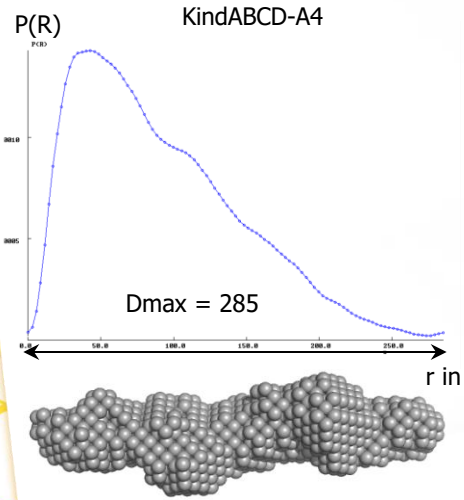
Distance Distribution Function

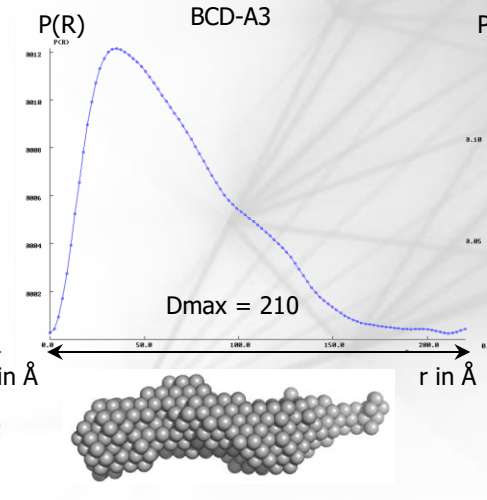
Scattering curves obtained on different complexes Spire-Actin and Actin alone

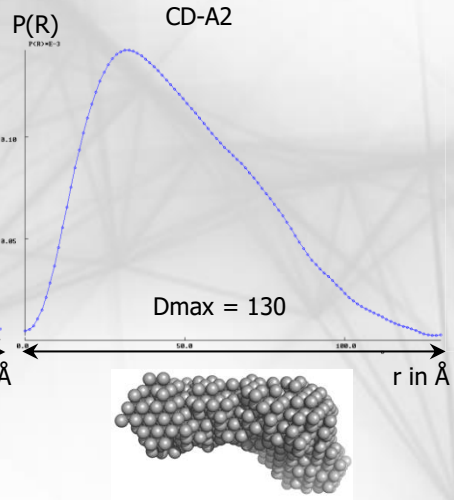


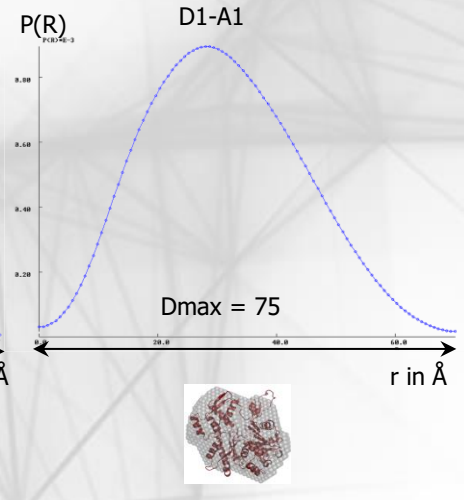
Complexes	Radius of gyration	Maximum diameter
	75.5 Å	285 Å
	55.5 Å	210 Å
	38.9 Å	130 Å
	25 Å	75 Å
	23.1 Å	70 Å

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









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)

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

First CONCLUSION

- ✓ SAXS is at his best when it is used to distinguish between several preconceived hypotheses.
- ✓ Analysis and modeling require a monodisperse and ideal solution, which has to be checked independently.
- ✓ Otherwise :

IN

SAXS

OUT

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)

Molecular modeling

Cristallographic , NMR structures or complete molecular modeling
- theoretical curves calculation and comparison with experimental data

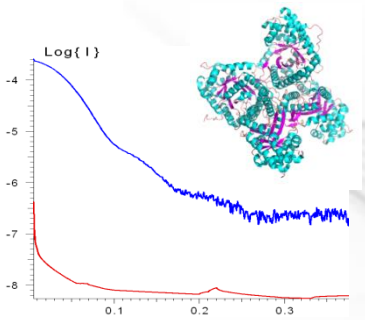
Nothing is known
- low resolution shape

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

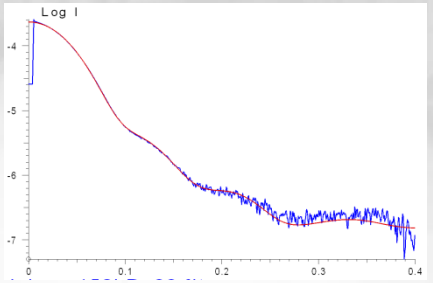
Structures with missing loop or flexible parts
- molecular modeling of missing parts against SAXS data

Modeling using SAXS data, available programs

1) Theoretical model or complete atomic structure available

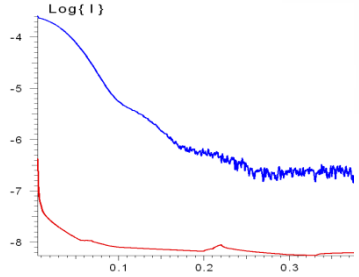


Validation/identification in solution

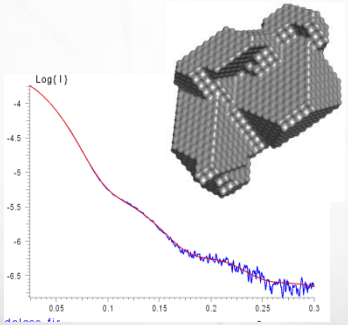


CRYSOL
FOXS

2) Nothing known (except the curve)

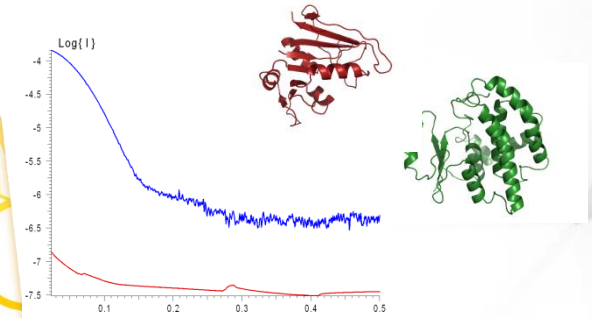


Low resolution model



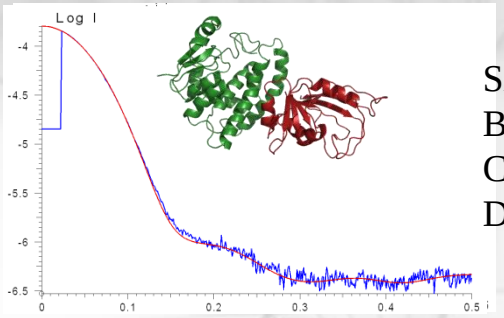
DAMMIN
DAMMIF
GASBOR
MONSA
DENFERT

3) Structures of subunits available



Rigid body modeling of the complex and

molecular modeling of the missing part



SASREF
BUNCH
CORAL
DADIMODO

Common features to all approaches

- ✓ *Monte-Carlo based methods (simulated annealing, genetic algorithm) : no unique solution.*

  ⟹ *repeat the calculation ca 10 times.*

  ⟹ *repeat the calculation n x 100 times followed by clustering.*

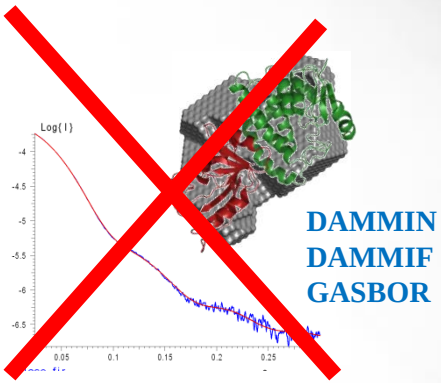
- ✓ *make use of constraints to restrict the solution space to (bio)physically meaningful models. The program minimizes the sum of the χ^2 with experimental data and penalty terms such as:*

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

A word of caution: what NOT to do

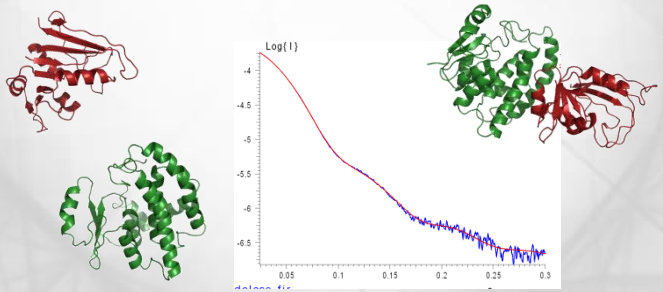
- **Common misconception: dummy atom ab initio envelope from DAMMIF (or from Gasbor for that matter) are viewed as similar to EM density maps: NO.**

- **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.**



DAMMIN
DAMMIF
GASBOR

SASREF

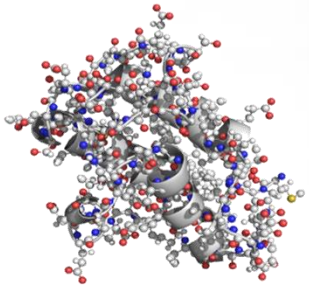


- **Furthermore, in some cases, the volume or envelope notion is simply irrelevant: for instance, for flexible multi domain proteins or even worse, for a flexible IDP.**

SAXS for 3D structure reconstitution

The 1D SAXS profile is the Fourier transform of the p(r) function. Contrary to direct scattering calculation, the inverse problem cannot be solved analytically, i.e., no "inverse computation" can be used to yield 3D position coordinates from scattering data.

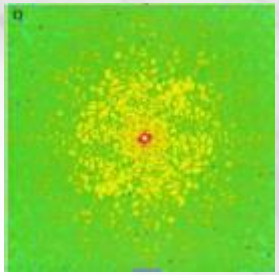
Real space 3D Molecule



Fixed orientation

Phase lost

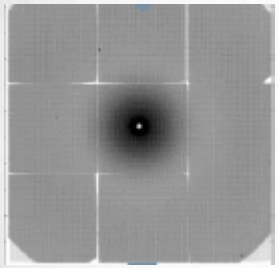
Reciprocal space
2D anisotropic image



Averaged orientation

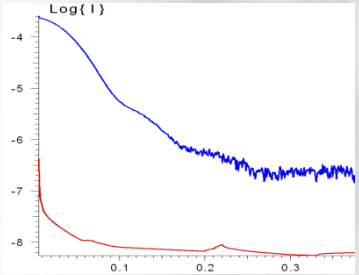
Orientation lost

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

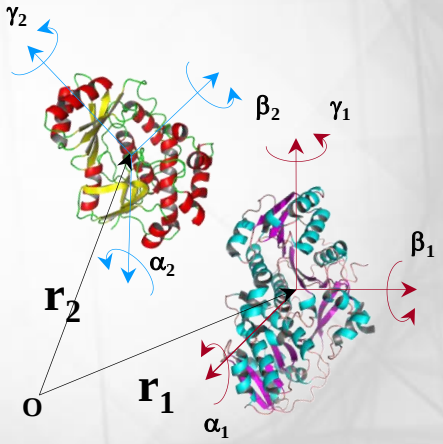
Rigid body modeling against SAXS data

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

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

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

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



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

The algorithm of minimization uses a penalty function (interconnectivity of the subunits, the steric clashes). It's possible 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.

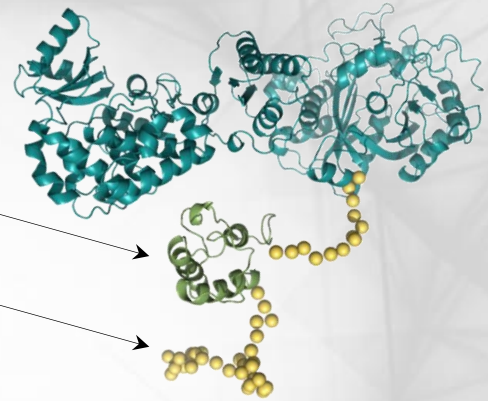
Rigid body modeling with missing loop against SAXS data

***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.



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

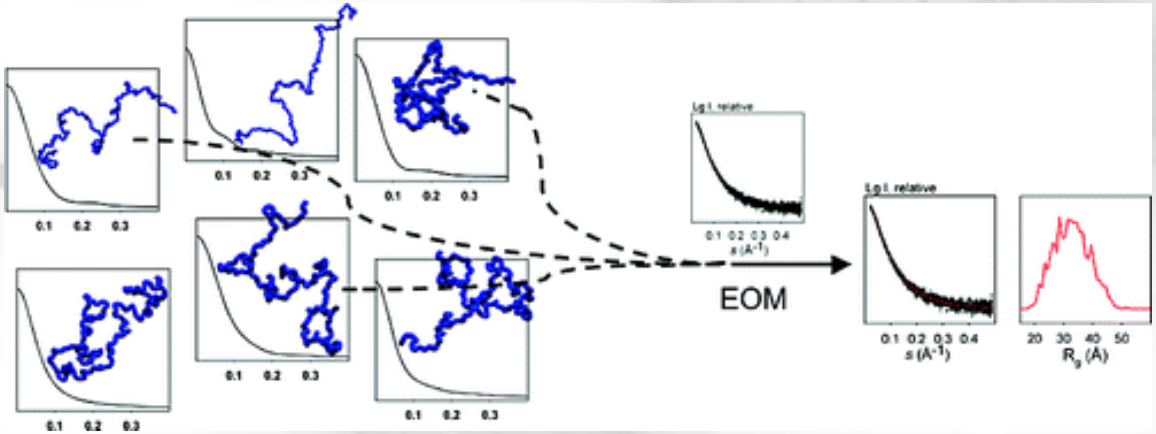
As in SASREF, amplitudes are calculated using CRYSOL from the high resolution structure of each domain.

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 experiments is also added.

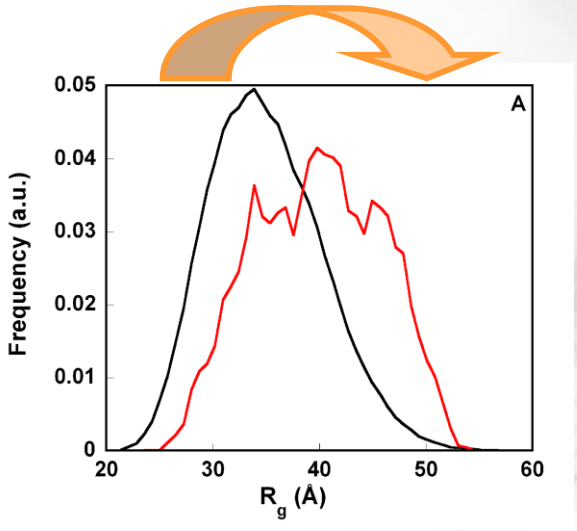
Flexibility → no unique structure !
NOT a structure but a SAXS data compatible model

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

Ensemble Optimized Method: EOM

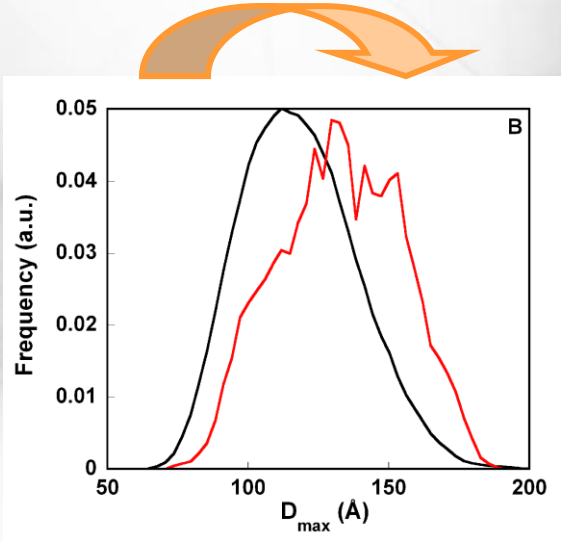


R_g distribution



— pool

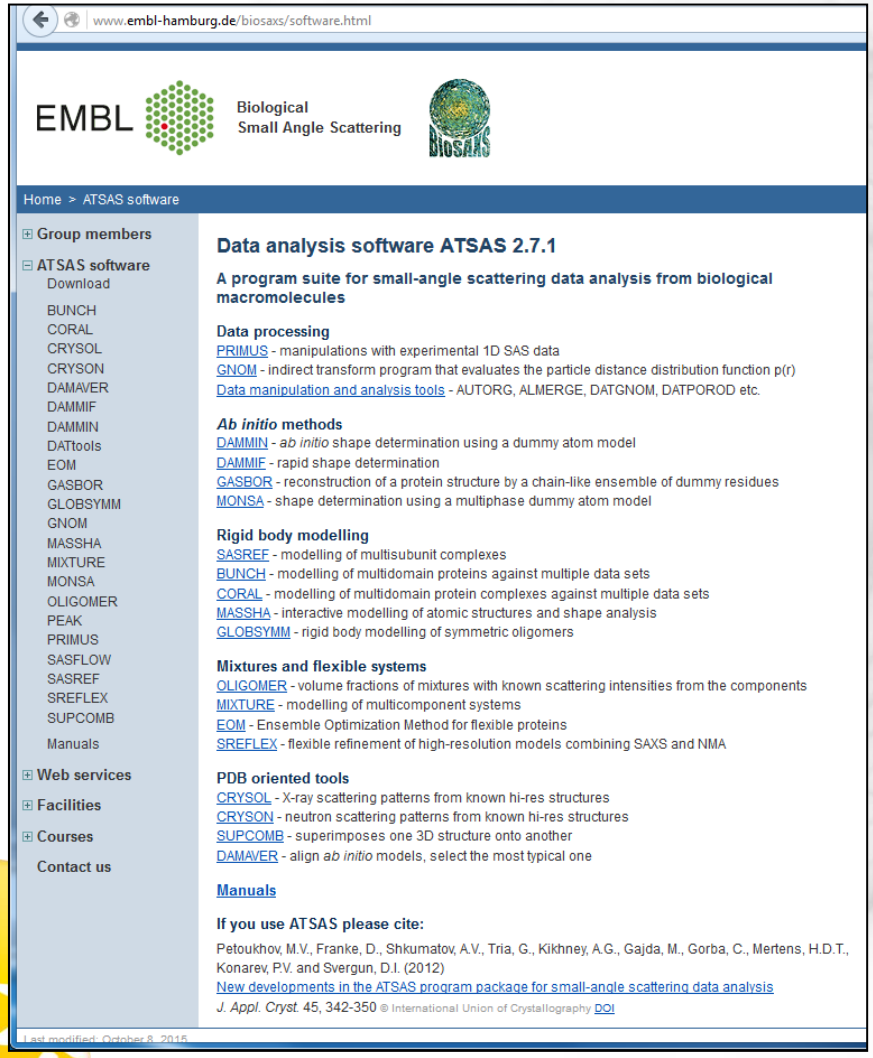
D_{Max} distribution



— optimized ensembles

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

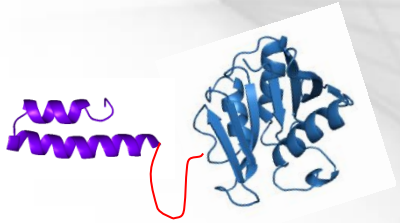
Collaboration : Christina Sizun & François Bontems (ICSN, Gif sur Yvette))

Evrard et al. (2011), *J. Appl. Cryst.*, **44**:1264-1271.

Modelling approach : complete atomic model

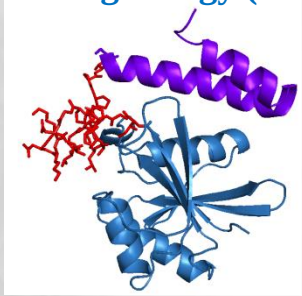
Full structure initiated with :

- Crystal or NMR domain structures
- Homology models



External information:

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

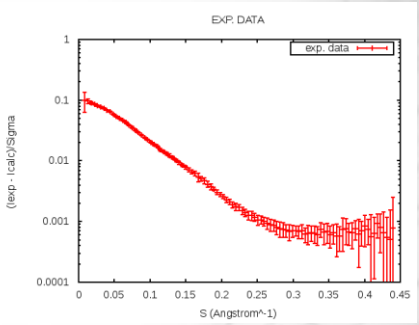


Experimental data:

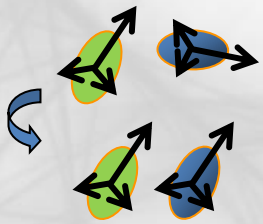
- SAXS
- NMR

  RDC

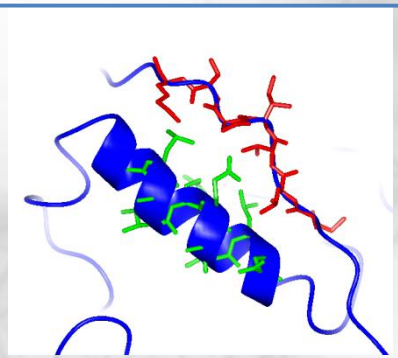
  chemical shift perturbations



SAXS score



RDC score



ADR score

Optimisation of the structure via a genetic algorithm

2nd CONCLUSION

- 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.