How to judge data quality

Tsutomu Matsui

SSRL Lab / Dept. of Chemistry
Stanford University

SLAC
NATIONAL ACCELERATOR LABORATORY

SSRL

SMB
Structural Molecular Biology

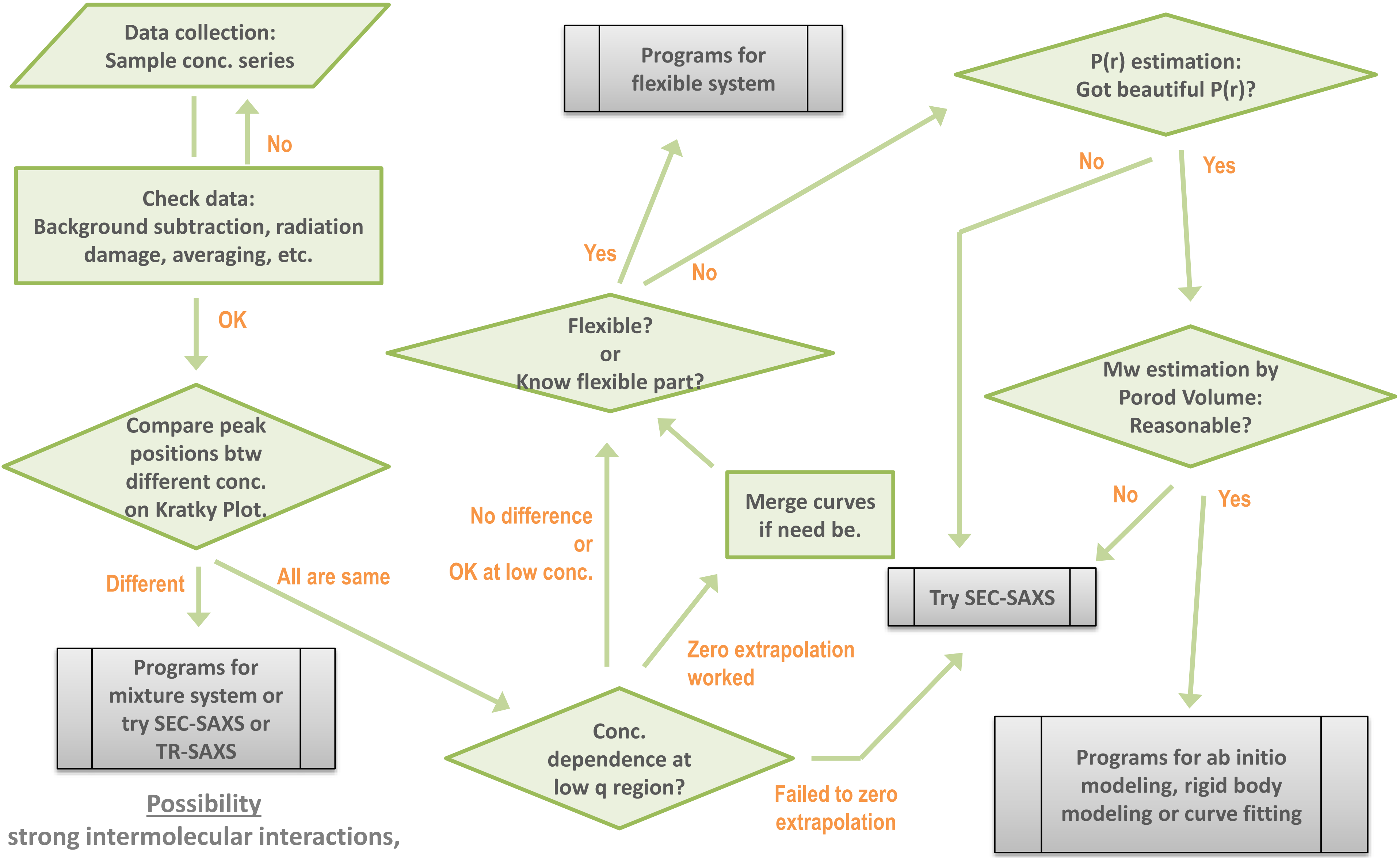
Subject of this session (Initial inspection of your data)

- You can feel “something wrong?” when something went wrong with your sample.
- You can assume what your SAXS curve is supposed to be.
- You can bring “true” SAXS data (scattering and/or P(r) profile) to further analysis.

**Regardless of whether the resulting structure is true or not,
the program “always” gives you structure.**

Note: In this session, unless otherwise noted, sample is assumed as monodisperse globular protein.

Basic Flow Chart



Data collection:
Sample conc. series

No

Check data:
Background subtraction, radiation damage, averaging, etc.

OK

Compare peak positions btw different conc. on Kratky Plot.

Different

Programs for mixture system or try SEC-SAXS or TR-SAXS

All are same

Conc. dependence at low q region?

No difference
or
OK at low conc.

Zero extrapolation worked

Merge curves if need be.

Failed to zero extrapolation

Flexible?
or
Know flexible part?

Yes

Programs for flexible system

No

P(r) estimation:
Got beautiful P(r)?

No

Yes

Mw estimation by Porod Volume:
Reasonable?

No

Try SEC-SAXS

Yes

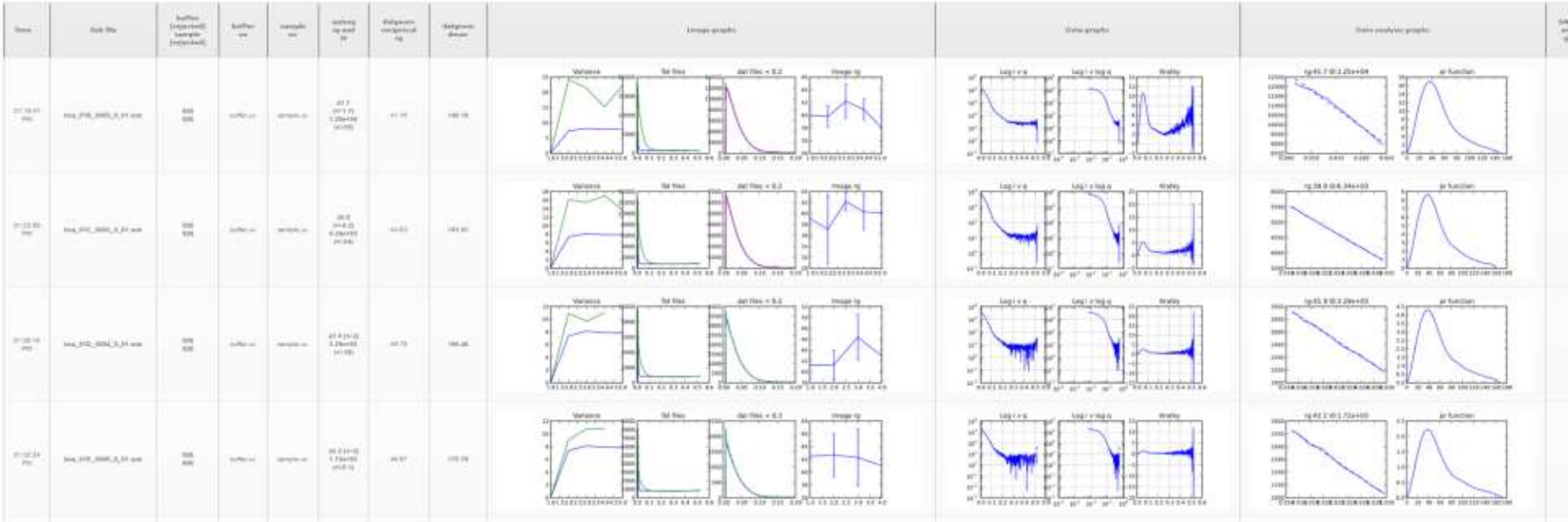
Programs for ab initio modeling, rigid body modeling or curve fitting

Possibility

strong intermolecular interactions, assembly, disassembly, dissociation, multimerization, etc.

SAXSPipe (automatic analysis pipeline at BL4-2)

- SAXSPipe is designed for a quick data check at beamline.
- Need to analyze your data carefully by hand!



Date

Data name

# of rejection file at averaging

Rg, I(0) and Dmax

Inspection of each frame

Rg of each frame

Semi-log plot

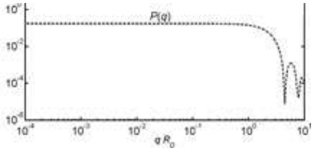
log-log plot

Kratky plot

Guinier plot

P(r)

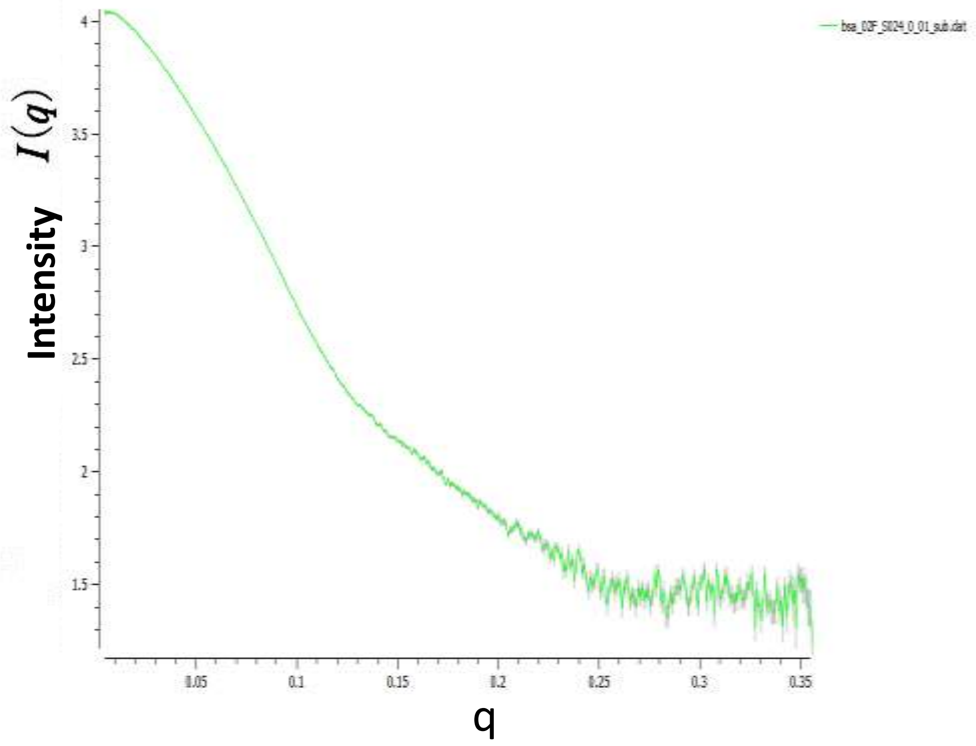
Monodisperse sample should have a plateau at low angle part of log-log plot.



Reference:
The accurate assessment of small-angle X-ray scattering data.
Grant TD, et, el., Acta Crystallogr D. 2015 71:45-56.

Shape of SAXS curve

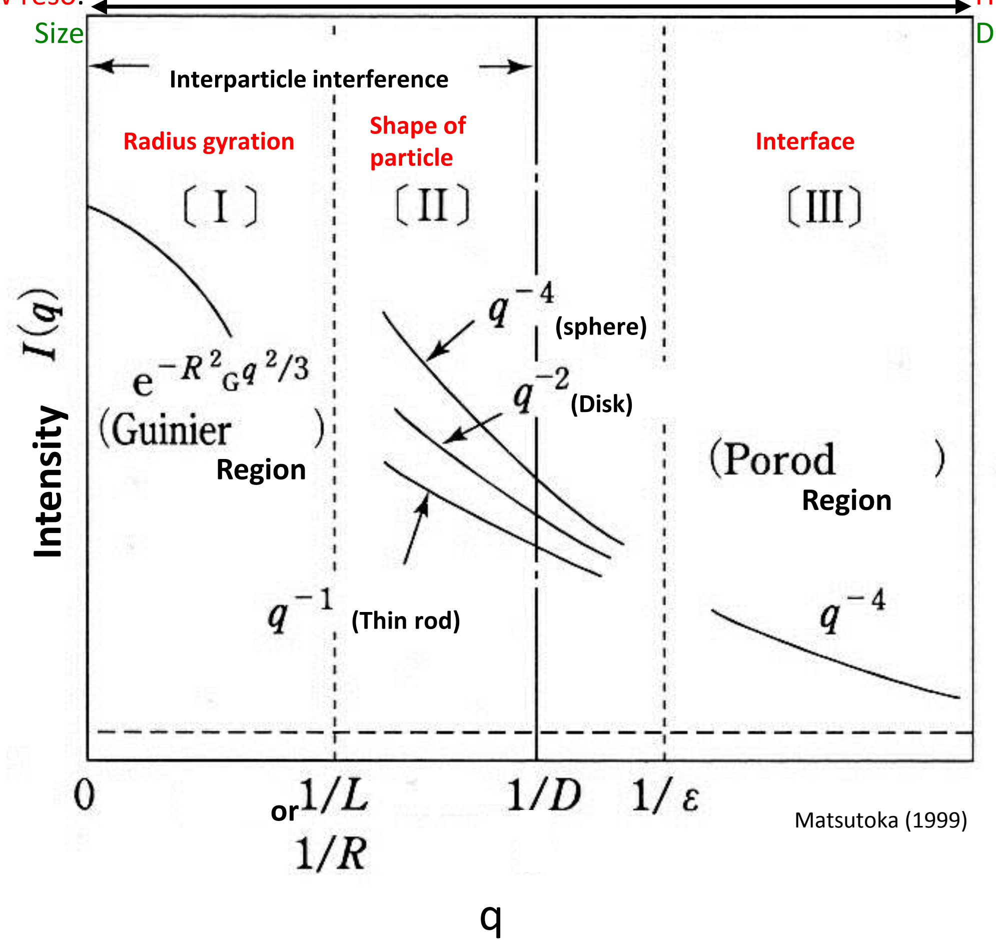
SAXS Data: Sample shape and intensity



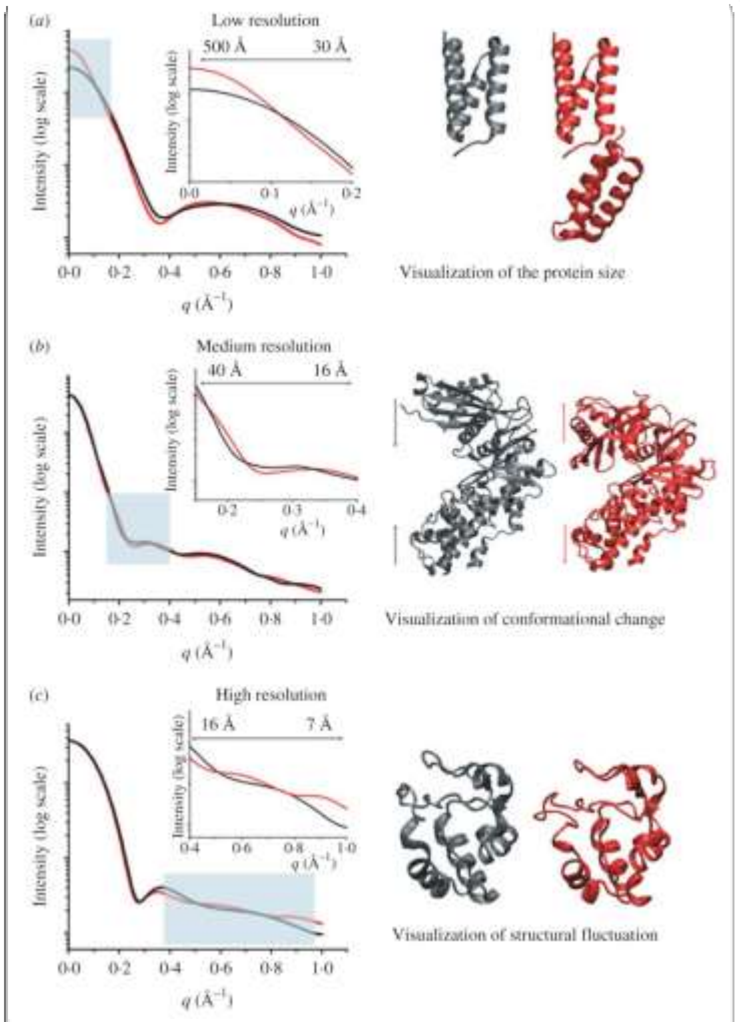
SAXS Data: Sample shape and intensity

Small angle
Low reso.
Size

W
Hi
De



Matsutoka (1999)



Putnam, et al. 2007 Quarterly Reviews of Biophysics 40, 191

Scattering profile of sphere

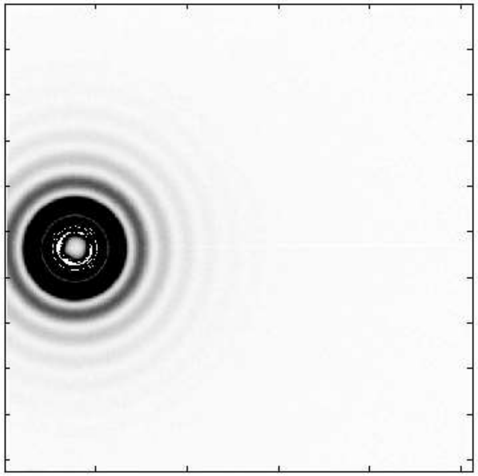
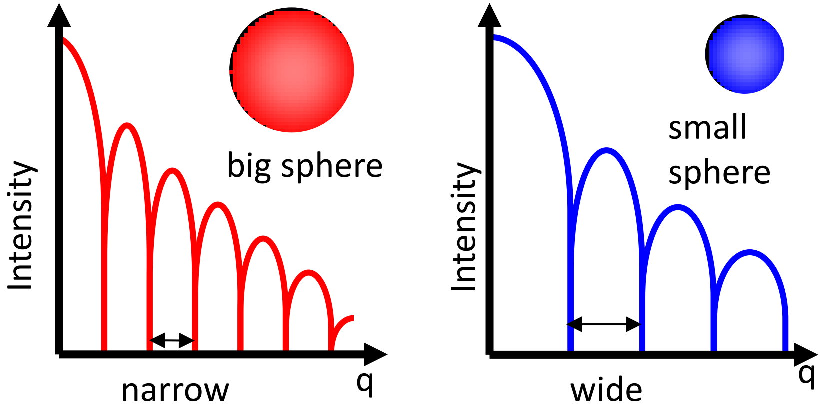
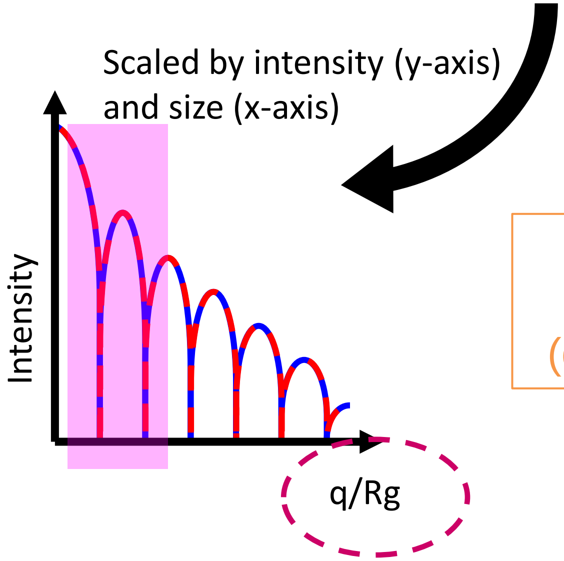


Image of silica spheres

$$I(q) = \rho_0^2 v^2 \frac{9(\sin qR - qR\cos qR)^2}{(qR)^6}$$



Hard to get high reso data from large sample.

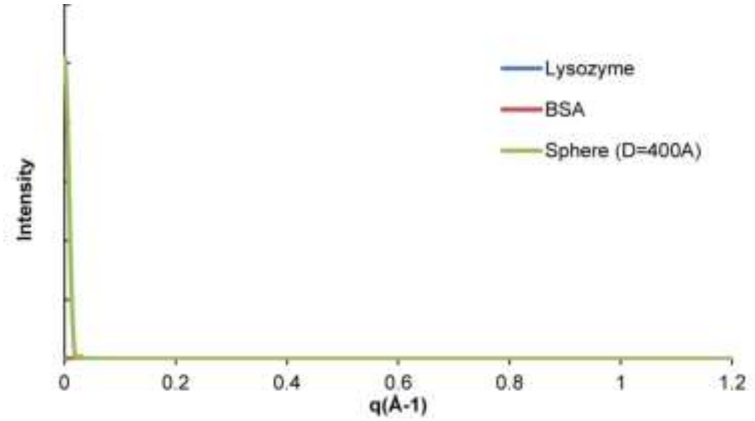


Scaled by intensity (y-axis) and size (x-axis)

Need right experimental setup depends on sample size. (detector distance & beam energy)

Sample size and intensity

- Theoretical SAXS profiles from a “globular” sample

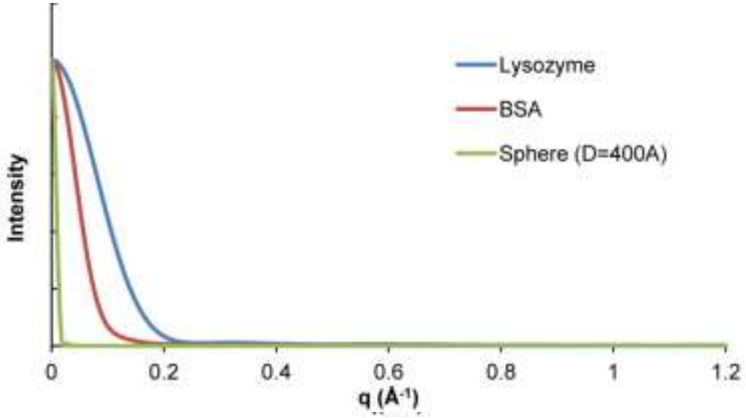


q-region containing size and shape information is size-dependent.
→ Need suitable experimental setup (e.g. x-ray energy, detector distance)

- Decay of intensity is associated with size.
- Guinier region is size-dependent.

Sample size and intensity

- Theoretical SAXS profiles from a “globular” sample

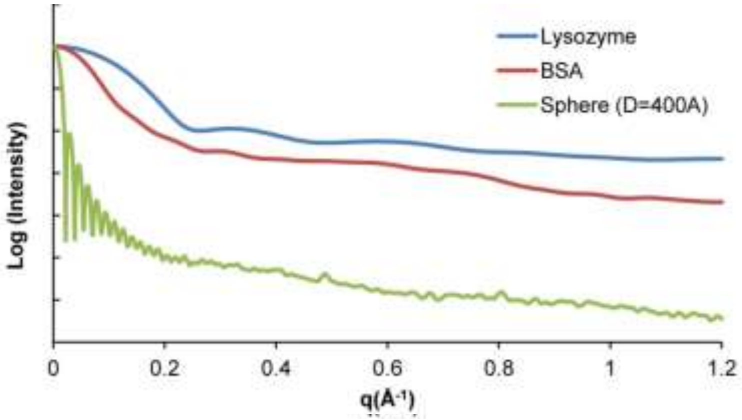


q-region containing size and shape information is size-dependent.
→ Need suitable experimental setup (e.g. x-ray energy, detector distance)

- Decay of intensity is associated with size.
- Guinier region is size-dependent.

Sample size and intensity

- Theoretical SAXS profiles from a “globular” sample

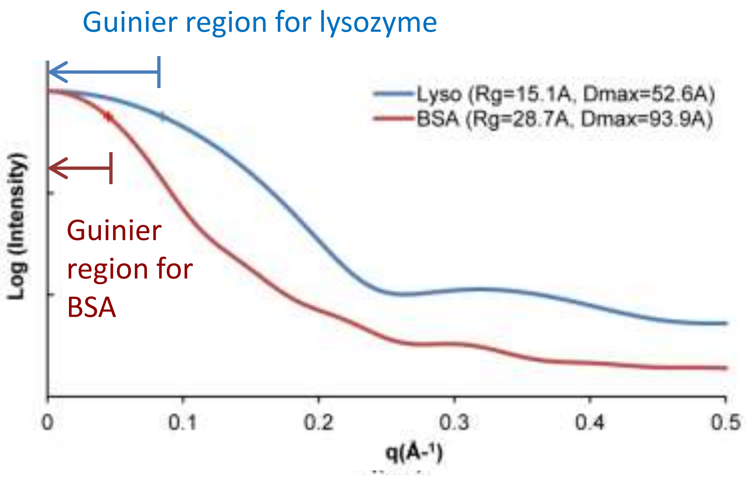


q-region containing size and shape information is size-dependent.
→ Need suitable experimental setup (e.g. x-ray energy, detector distance)

- Decay of intensity is associated with size.
- Guinier region is size-dependent.

Sample size and intensity

- **Theoretical SAXS profiles from a “globular” sample**

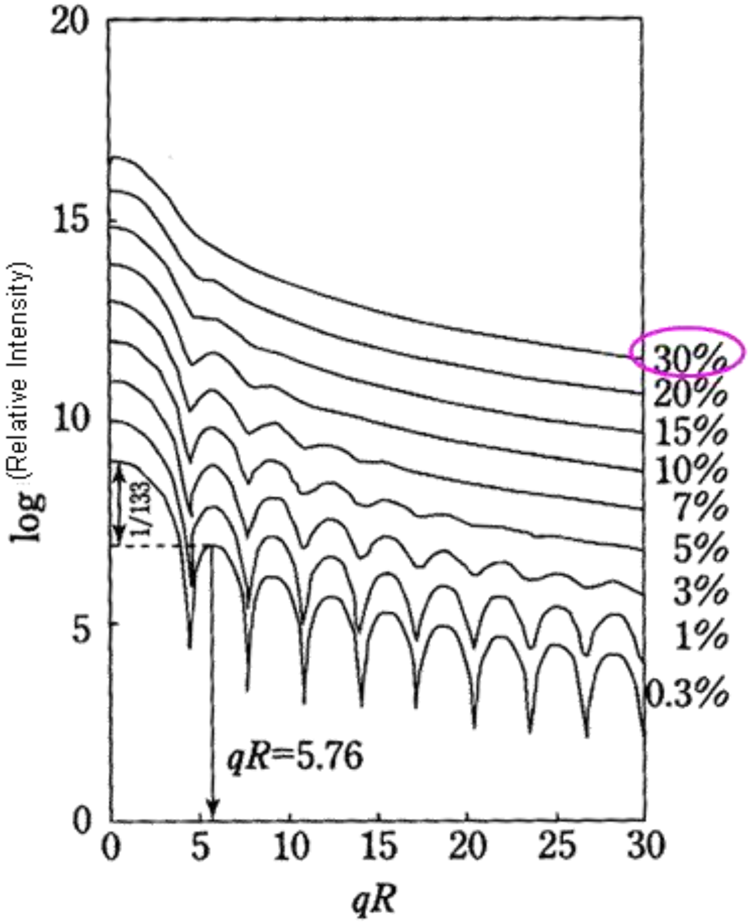


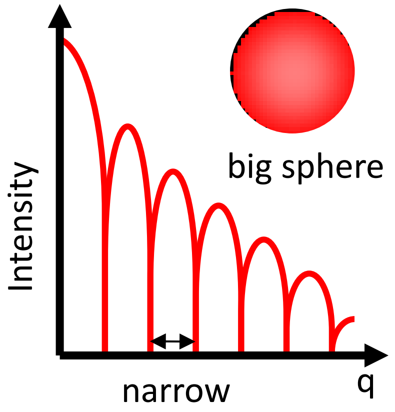
q-region containing size and shape information is size-dependent.
→ Need suitable experimental setup (e.g. x-ray energy, detector distance)

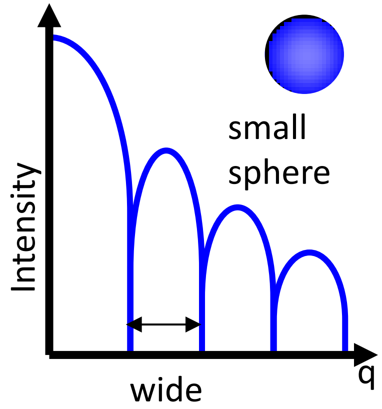
- Decay of intensity is associated with size.
- Guinier region is size-dependent.

Polydisperse sample

- SAXS curves of sphere at different polydispersity



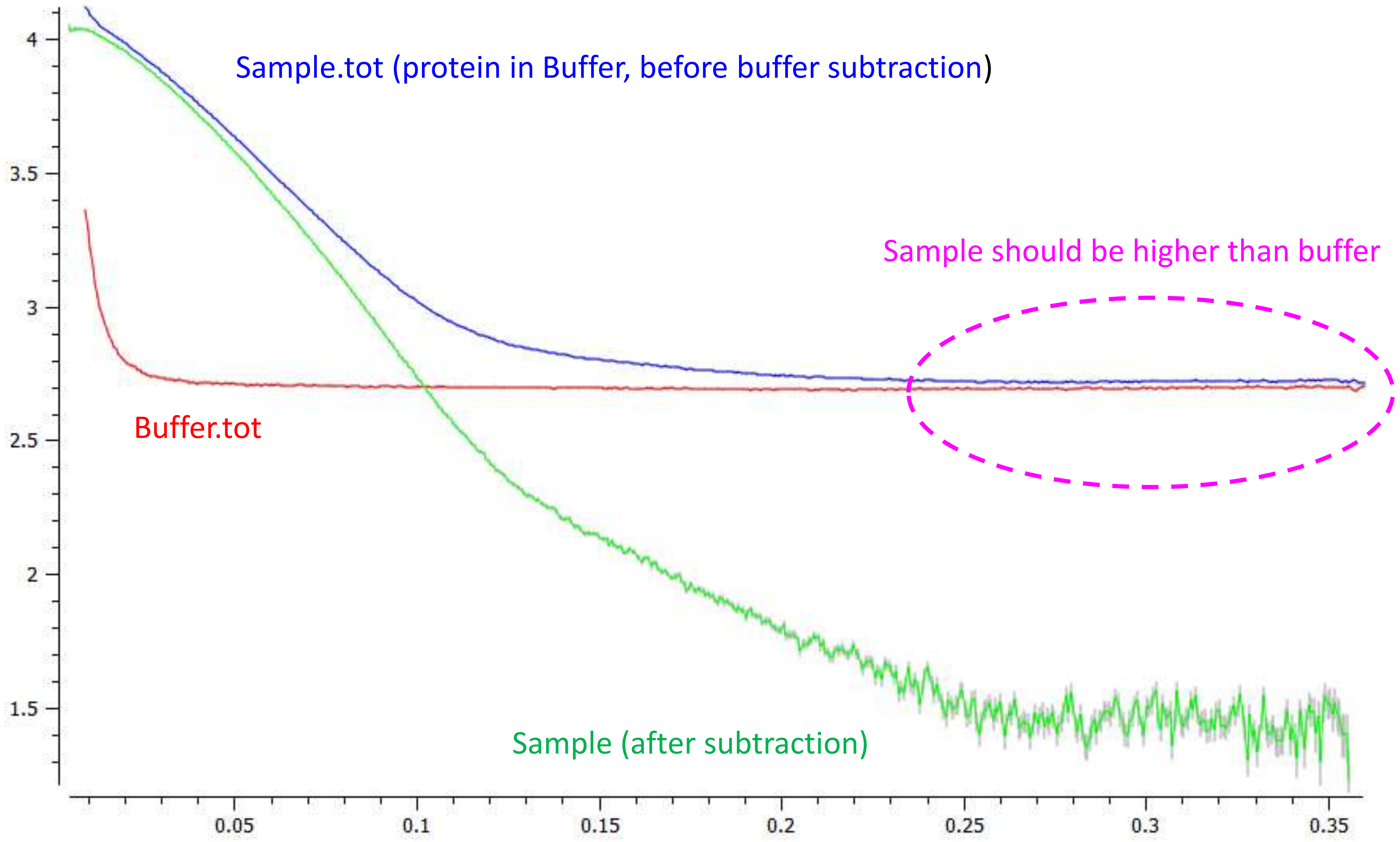




Background subtraction, radiation damage & concentration-dependence

Inspection after taking datasets of concentration series

Background subtraction

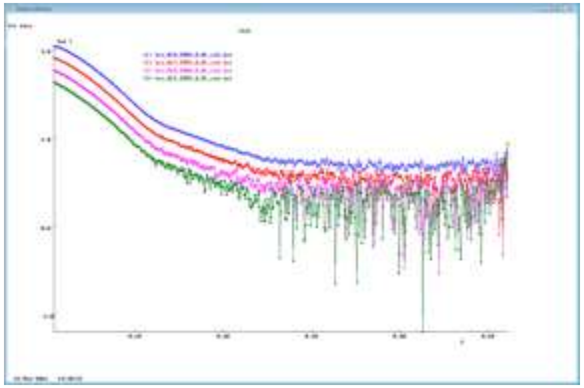


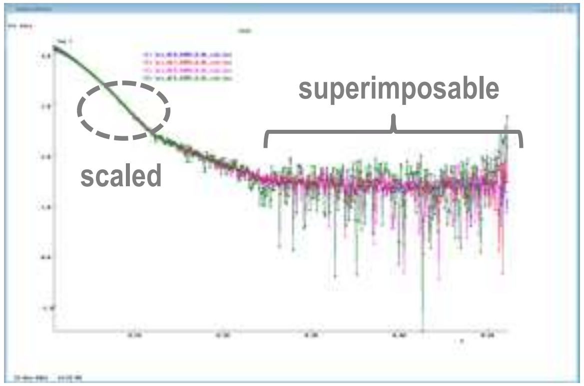
- **Need to double-check sastool output files in your lab (especially tot files)!**

Over and under subtraction

Right buffer

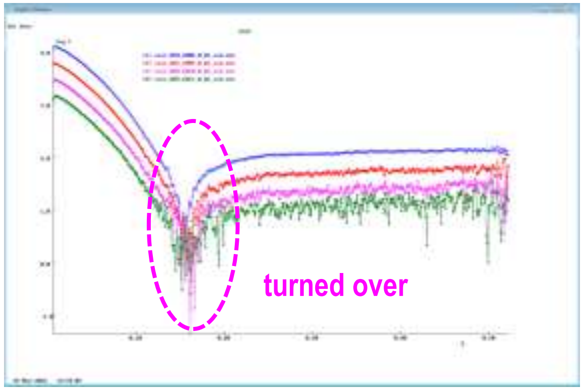
Blank buffer is identical to that of protein solution. All curves (different concentrations) have same level at high q.

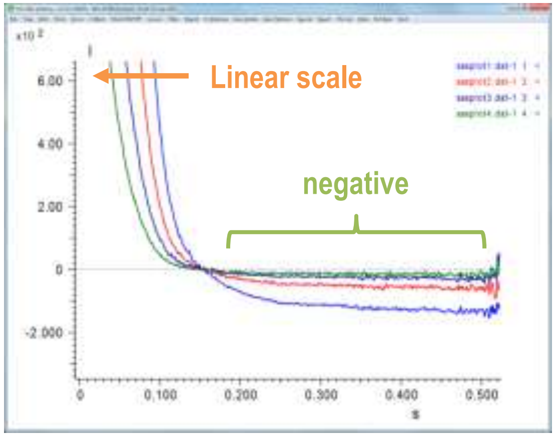




Over subtraction

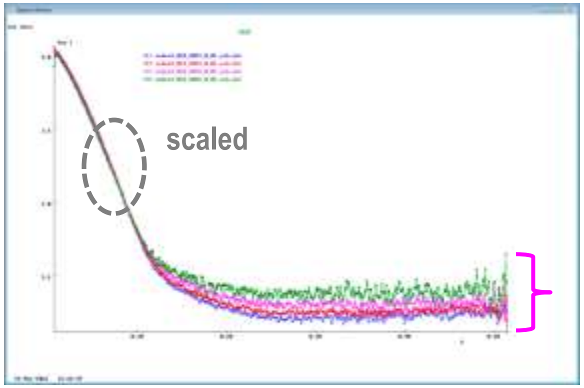
Ex) Only buffer contains 10% glycerol. Dilution series were made by the buffer without glycerol.





Under subtraction

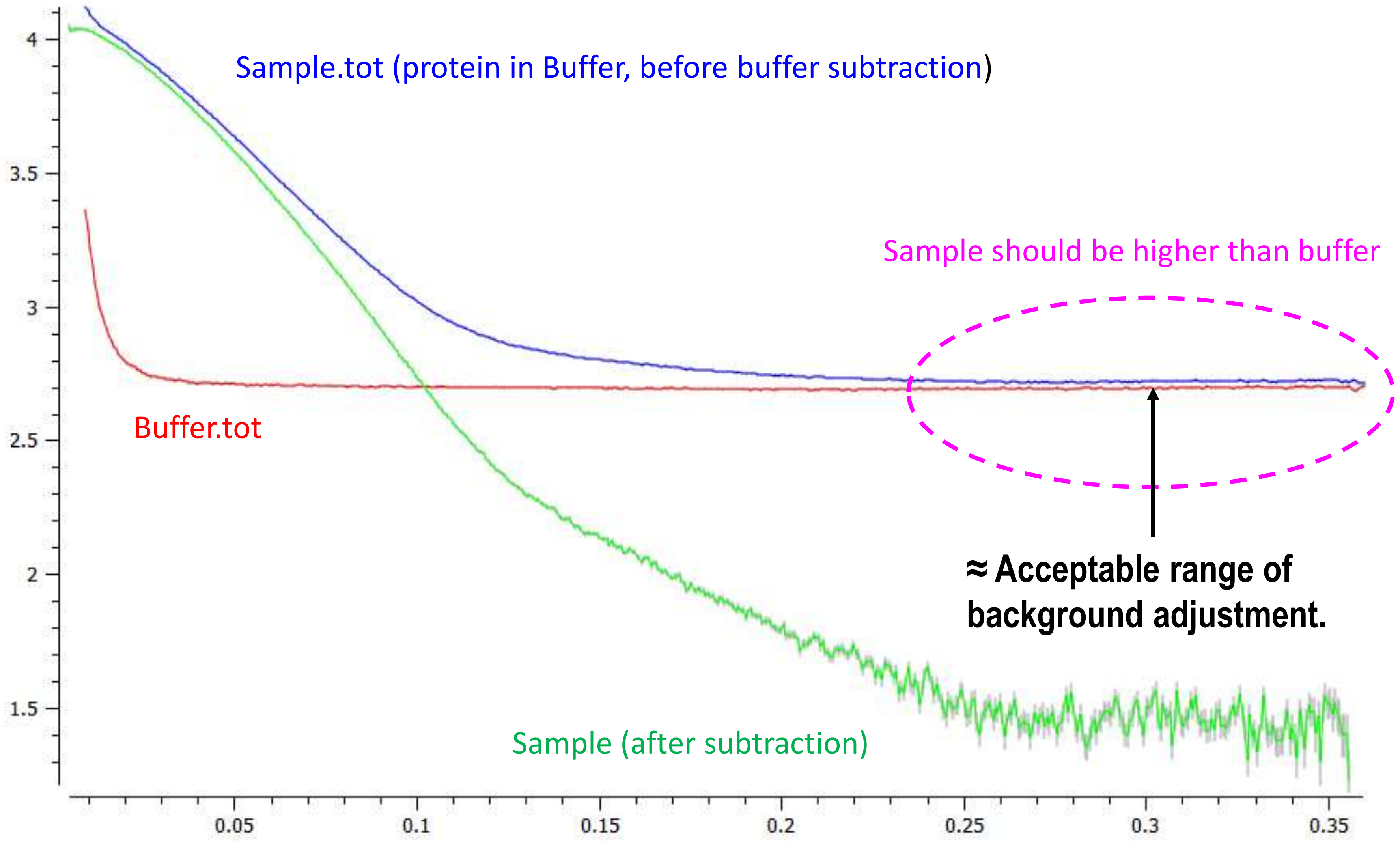
Ex) Buffer mismatch between protein solution, blank buffer and buffer used for making dilution series (10%, 0% and 1% glycerol, respectively).



Ordinary case

- Insufficient buffer exchange
- Glycerol used for storage
- Scaling error due to very weak signal

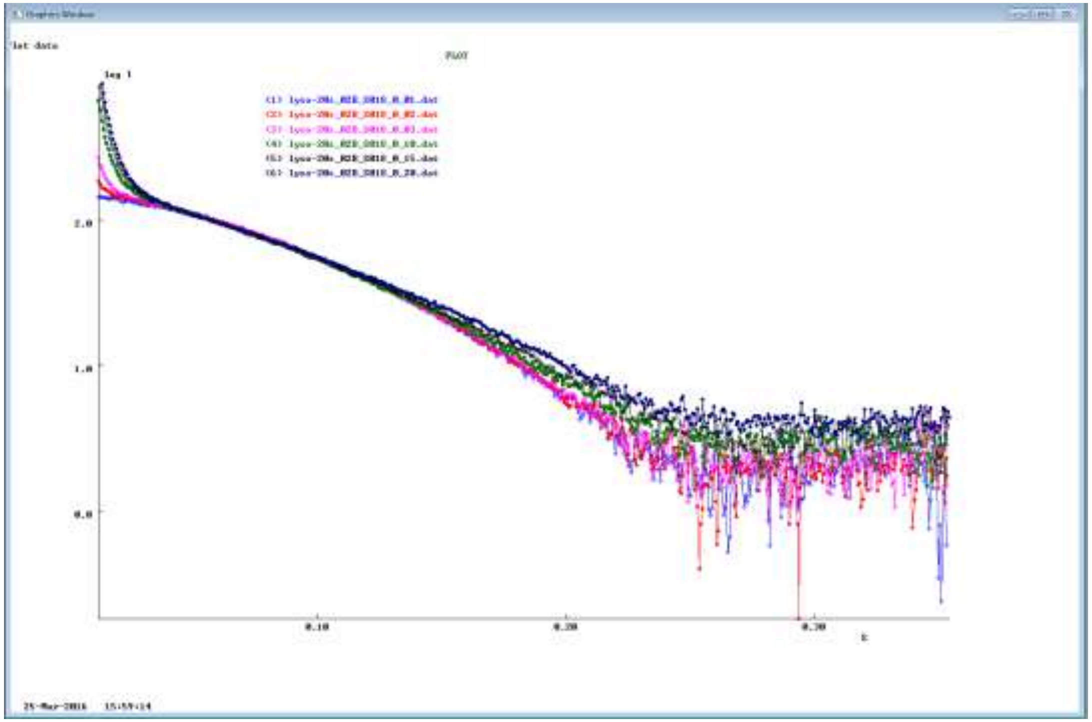
Background subtraction



- Need to double-check sastool output files in your lab (especially tot files)!

Radiation damage

- Lysozyme: 1sec x 20 images. No DTT in buffer.



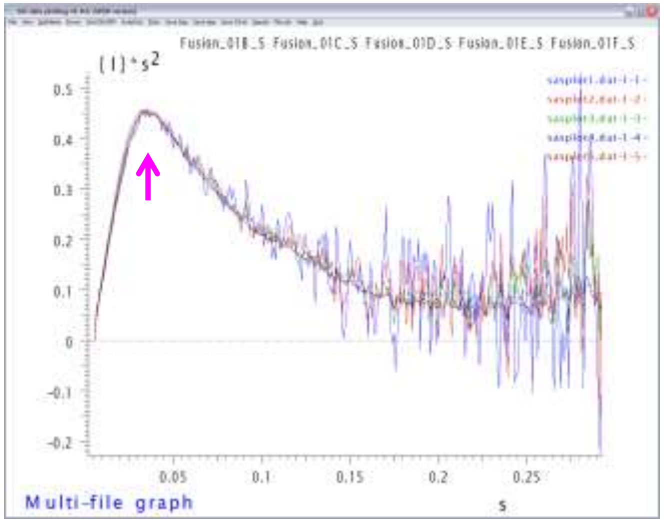
- Lysozyme with multilayer beam (flux: approx. x30 higher)

- Radiation damaged sample frequently adheres to capillary cell.
- Add radical scavenger (e.g. 5mM DTT) in sample and buffer.

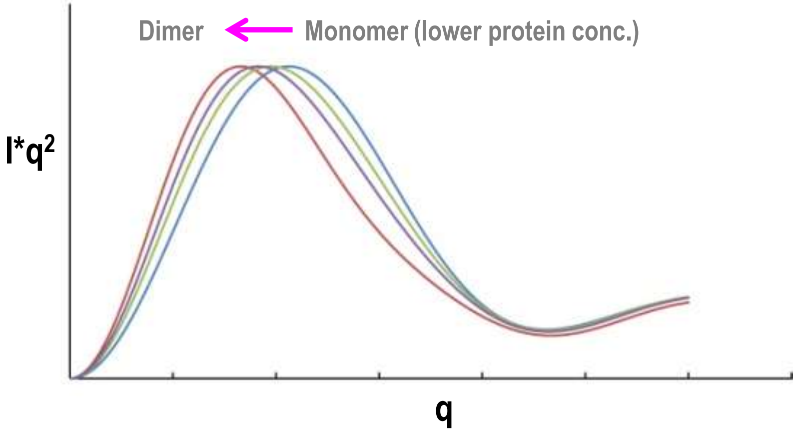
Check for all sizes of scatterer at different concentrations

- **Peak position on Kratky plot is associated with size of scatterer (sample).**
  - Scale all curves (different concentrations) at peak position on Kratky plot.
  - Peak position is generally insensitive to weak interparticle (intermolecular) interactions.
  - Be very careful about father analysis if peak positions are different.
  - Do not merge curves if peak positions are different.

Different protein concentration curves on Kratky plot



Monomer-Dimer equilibrium



Similar phenomenon could be observed with: Strong intermolecular interactions, assembly, disassembly, dissociation from complex, multimerization, aggregations, etc.

Pertinent details: Dimentionless Kratky plot

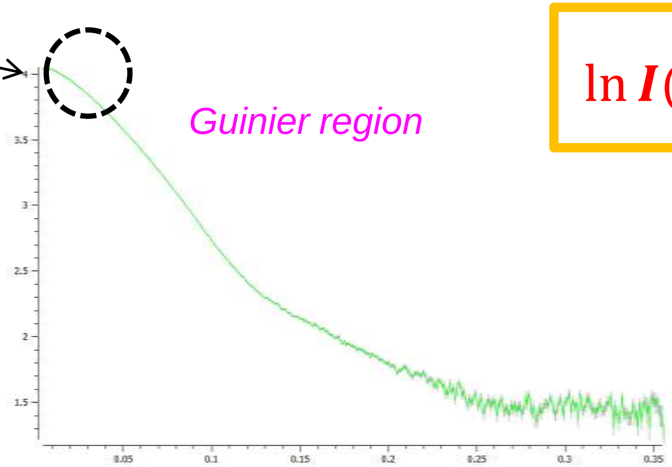
Guinier Analysis

Estimation of radius of gyration (R_g) and I(0)

$$I(0) = N(\Delta\rho V)^2 = C\Delta\rho^2 v^2 Mw/N_A$$

$\Delta\rho$: contrast

V : partial specific volume



Guinier region

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

Guinier Analysis

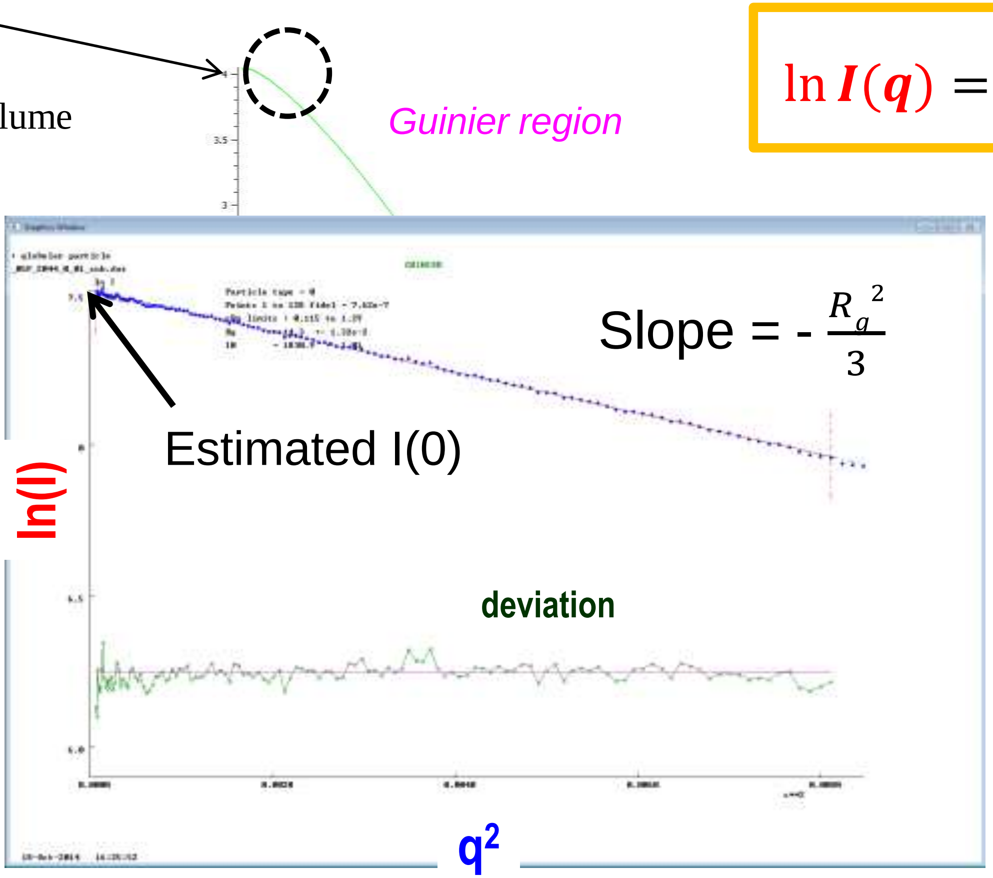
Estimation of radius of gyration (R_g) and I(0)

$$I(0) = N(\Delta\rho V)^2 = C\Delta\rho^2 v^2 Mw/N_A$$

$\Delta\rho$: contrast

V : partial specific volume

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



Guinier region

Slope = $-\frac{R_g^2}{3}$

Estimated I(0)

ln(I)

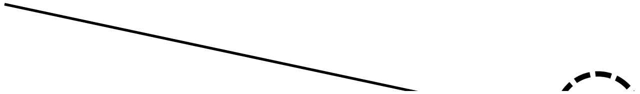
deviation

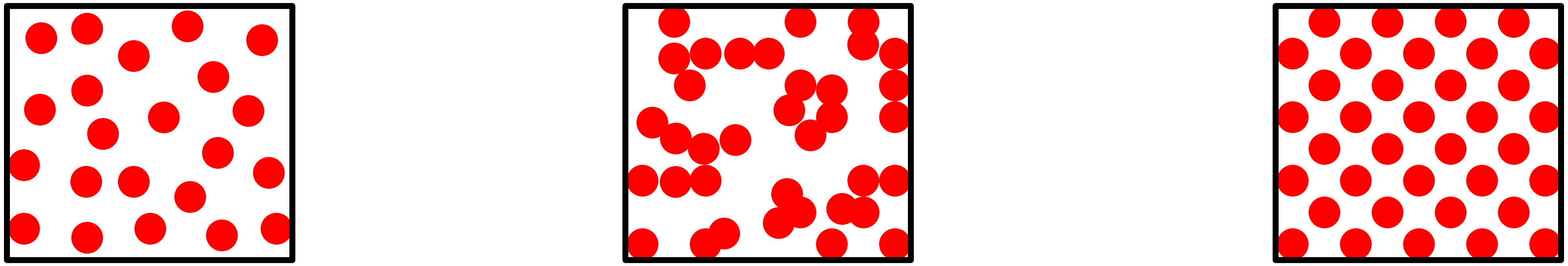
q^2

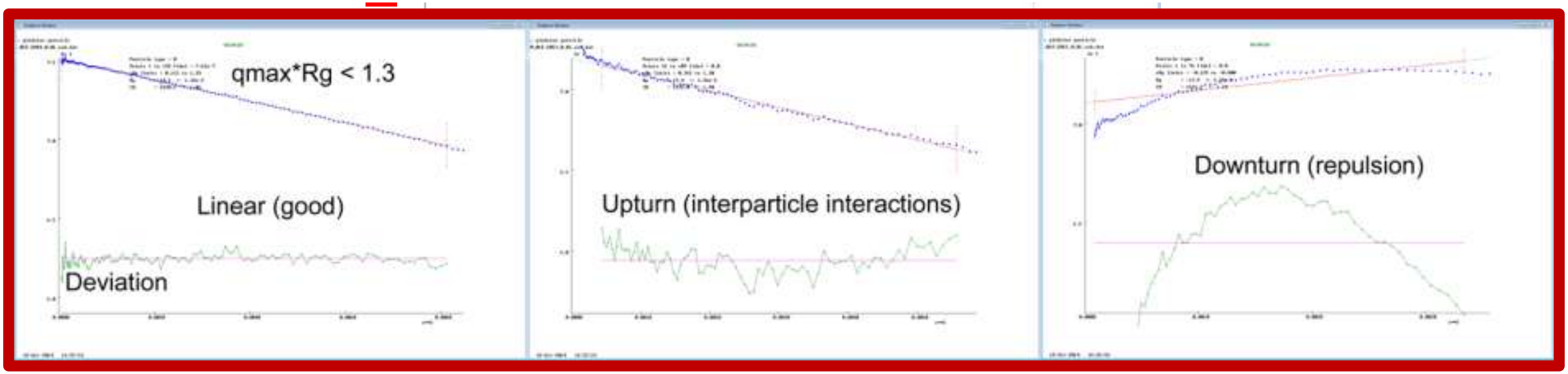
Guinier Analysis

Estimation of radius of gyration (R_g) and I(0)

$$I(0) = N(\Delta\rho V)^2 = C\Delta\rho^2 v^2 Mw/N_A$$







qmax*Rg < 1.3

Linear (good)

Deviation

Upturn (interparticle interactions)

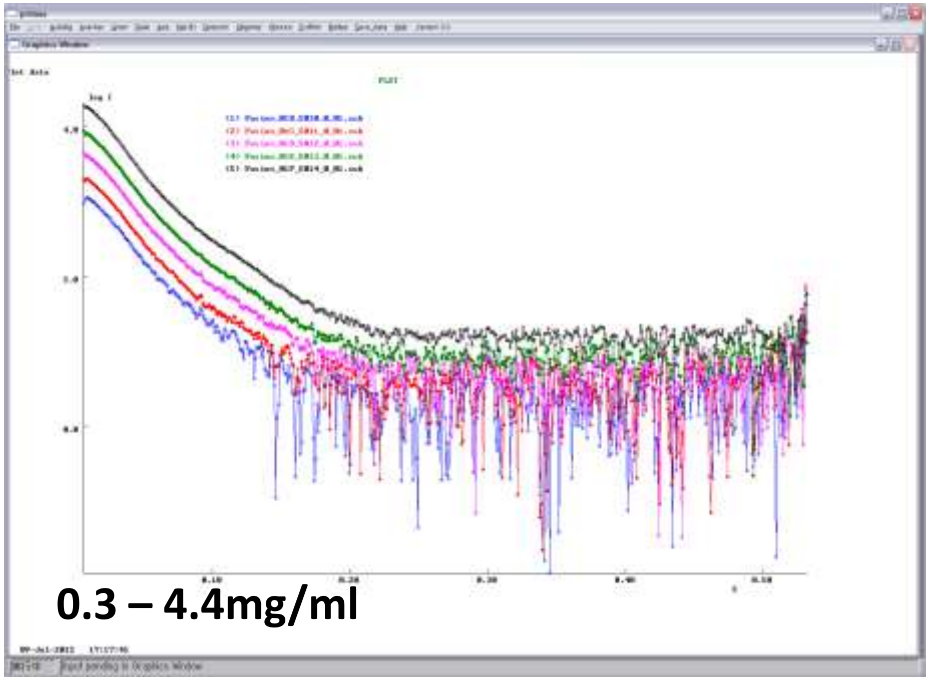
Downturn (repulsion)

Concentration dependence

$I(q) = P(q)\ S(q)$

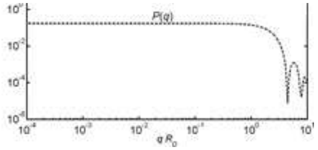
I(q): Intensity
P(q): Form factor of your sample
S(q): Structural factor

→ $I(q) = I_{monomer}(q)\ S_{intermolecular}(q)$



0.3 – 4.4mg/ml

log-log plot



Monodisperse sample should have a plateau at low angle part of log-log plot.

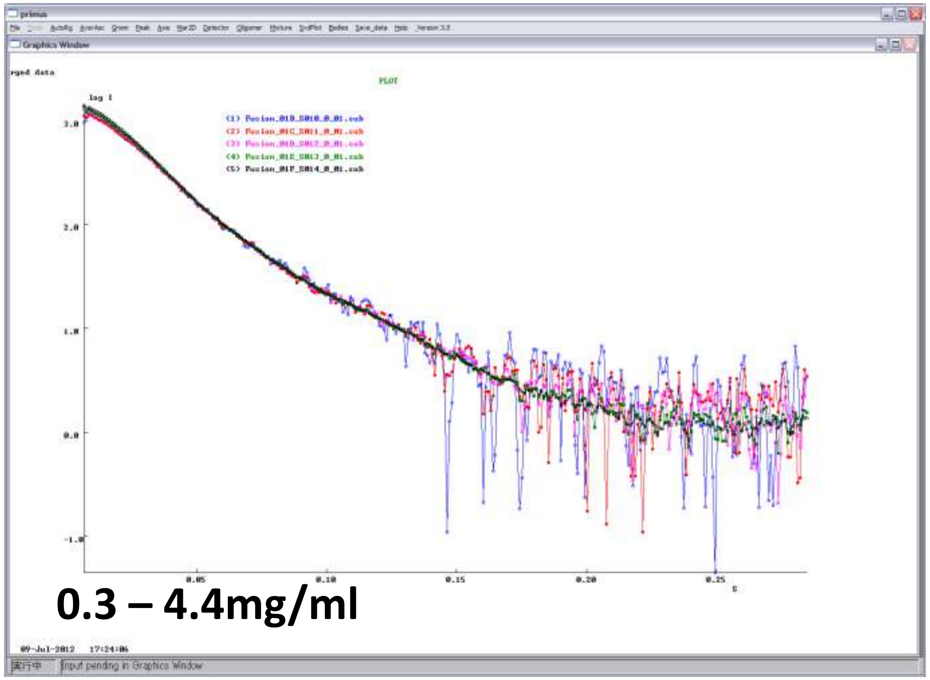
- Ideally, concentration-dependence could be negligible at lower concentrations (no difference at a few low concentrations). Otherwise zero-extrapolation is required (see next).
- Need to measure sample as low a concentration as experimentally possible.
- Buffer is reasonable?

Concentration dependence

$I(q) = P(q)\,S(q)$

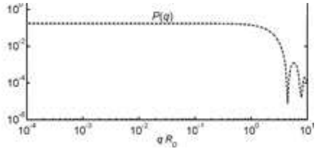
I(q): Intensity
P(q): Form factor of your sample
S(q): Structural factor

$\Rightarrow$ $I(q) = I_{monomer}(q)\,S_{intermolecular}(q)$



0.3 – 4.4mg/ml

log-log plot



Monodisperse sample should have a plateau at low angle part of log-log plot.

- Ideally, concentration-dependence could be negligible at lower concentrations (no difference at a few low concentrations). Otherwise zero-extrapolation is required (see next).
- Need to measure sample as low a concentration as experimentally possible.
- Buffer is reasonable?

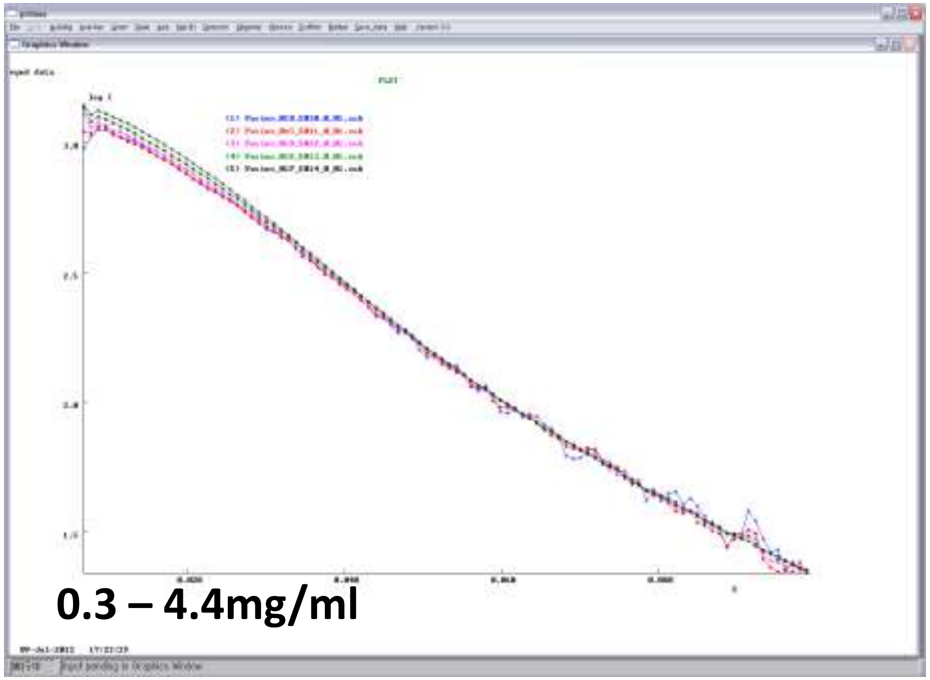
Concentration dependence

$I(q) = P(q)\ S(q)$

I(q): Intensity
P(q): Form factor of your sample
S(q): Structural factor

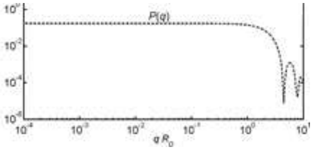
➡

$I(q) = I_{monomer}(q)\ S_{intermolecular}(q)$



0.3 – 4.4mg/ml

log-log plot

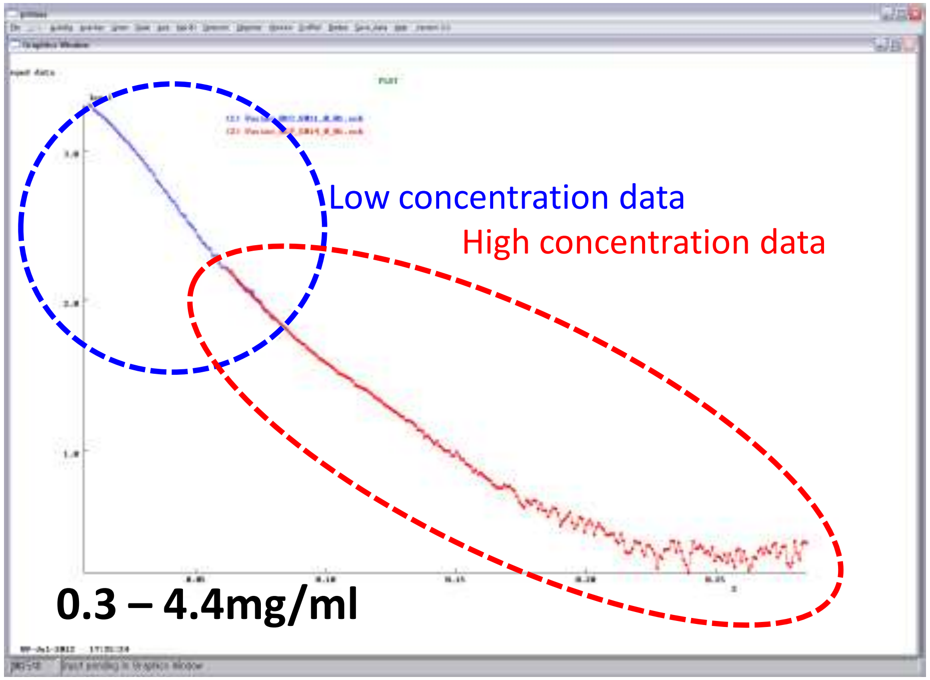


Monodisperse sample should have a plateau at low angle part of log-log plot.

- Ideally, concentration-dependence could be negligible at lower concentrations (no difference at a few low concentrations). Otherwise zero-extrapolation is required (see next).
- Need to measure sample as low a concentration as experimentally possible.
- Buffer is reasonable?

Merge curves

- Merging curves manually



Low concentration data

High concentration data

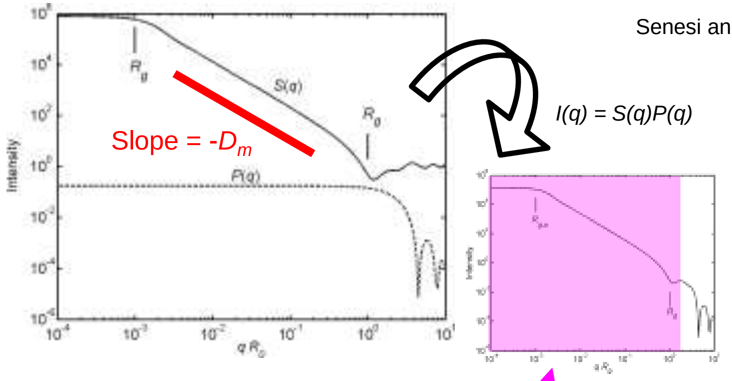
0.3 – 4.4mg/ml

PRIMUS (ATSAS)

- Automatic merge programs are available.
  - Almerge (ATSAS)
  - SAXS Merge (UCSF)
  - etc.

Aggregates

- **Aggregate could disturb your data depending on its size and density.**



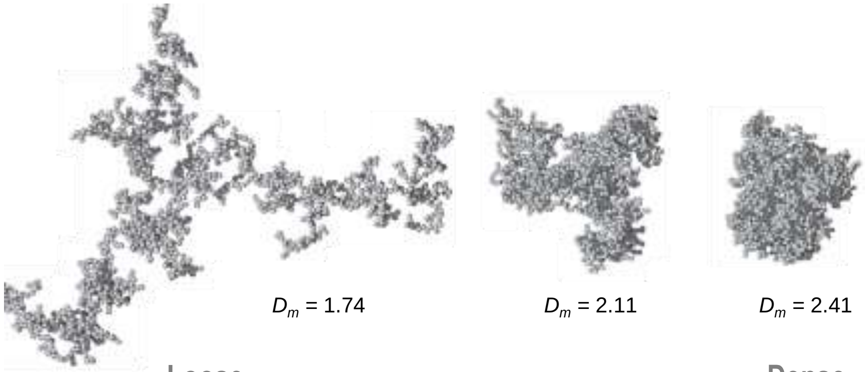
Senesi and Wilkinson (2008)

$I(q) = S(q)P(q)$

Slope = $-D_m$

At q << $1/R_0$ on a log-log plot, slope of Structural factor, S(q), is corresponding to mass fractal dimension (D_m), whereas Form factor of monodisperse sphere, P(q), is consistent at the region.

Simulation of fractal aggregates with 2000 spheres

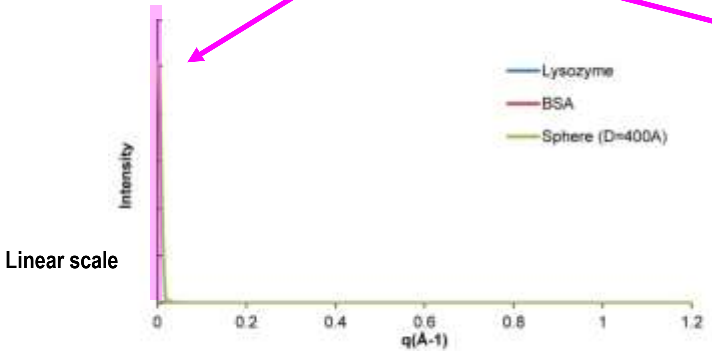


$D_m = 1.74$ $D_m = 2.11$ $D_m = 2.41$

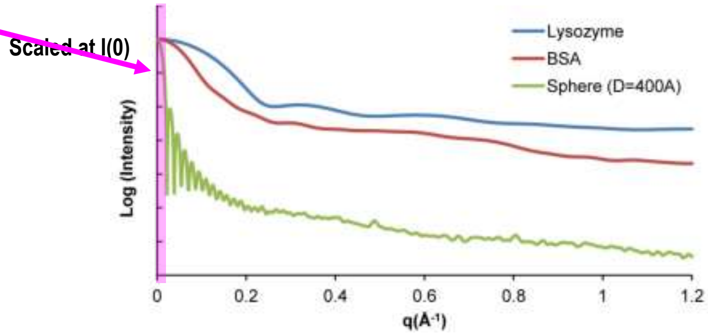
Loose aggregates ⟷ Dense aggregates

Beamstop

Scattering from a globular sample

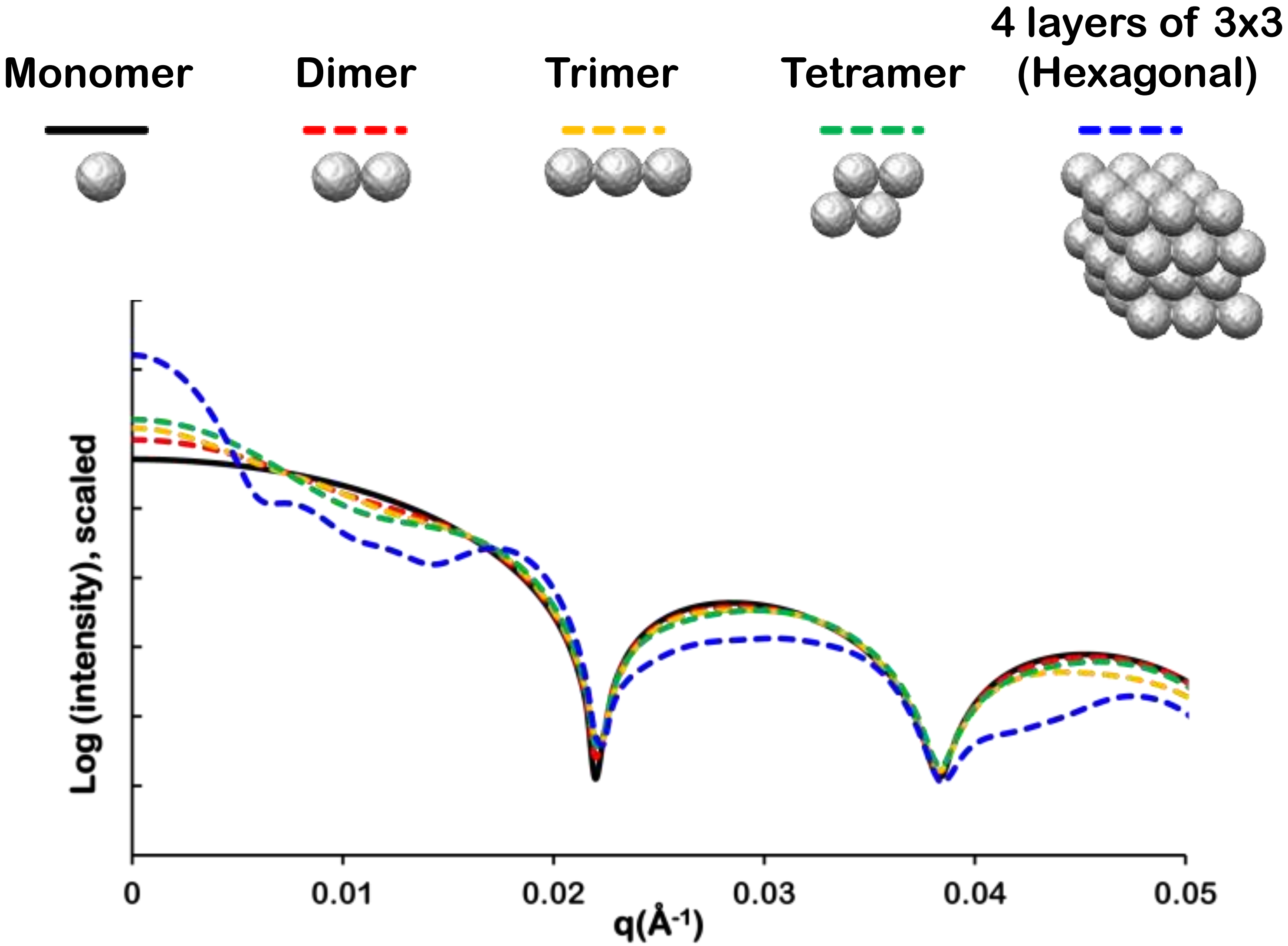


Linear scale



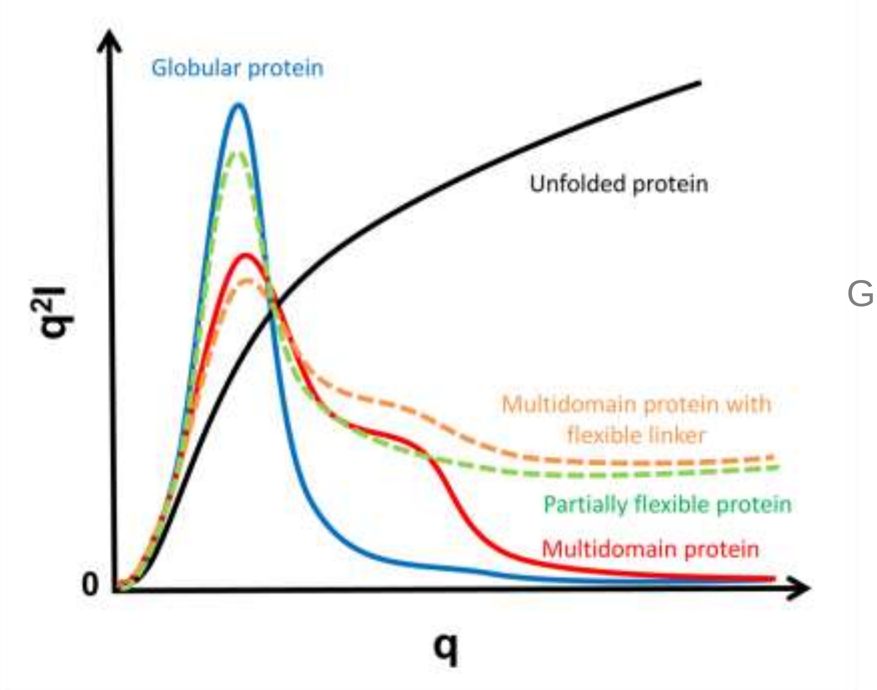
Scaled at I(0)

Effect of close-packing (R=200Å)



Kratly plot

- Good tool to evaluate flexibility and morphology of your sample.
- Compute theoretical curve if atomic structure or model is available.
- Sample has flexible loop or tag?
- Need to choose right software/program for further analysis.



Power law

Globular: q^{-4}

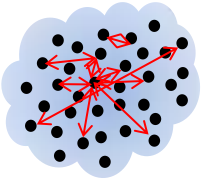
Rod:q^{-1}

Disk:q^{-2}

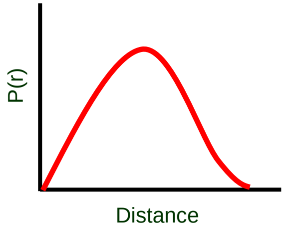
Gaussian chain: q^{-2}

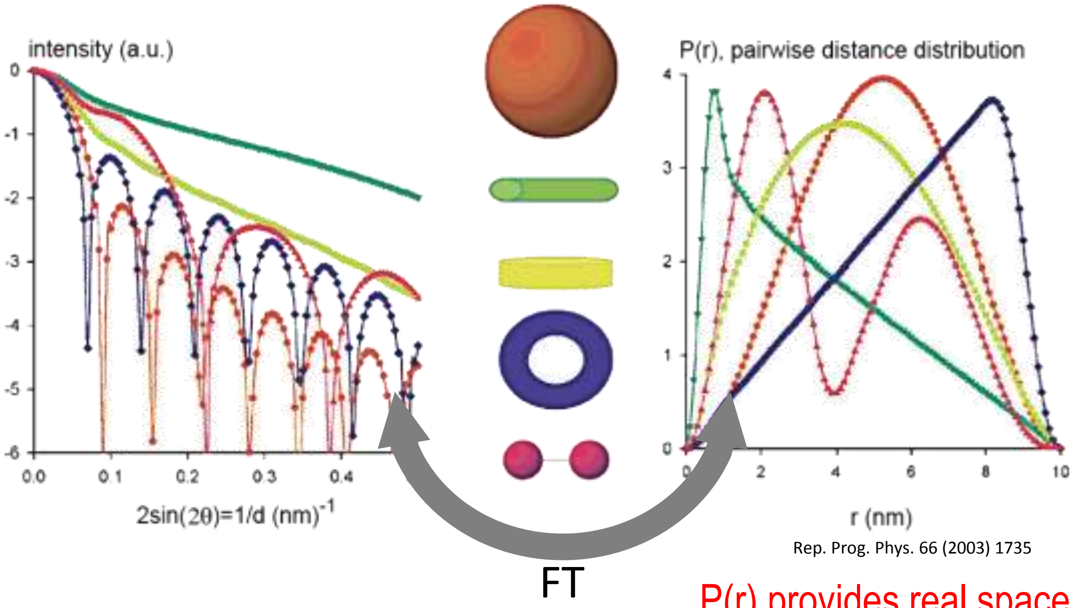
P(r), Pair-wise distance distribution function (PDDF)

- P(r) is inter-atomic distance distribution.



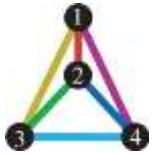
Histogram Of
inter-atomic distances



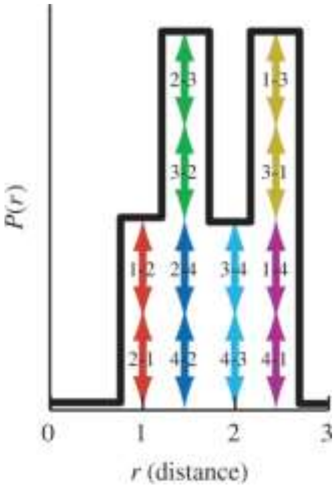


Rep. Prog. Phys. 66 (2003) 1735

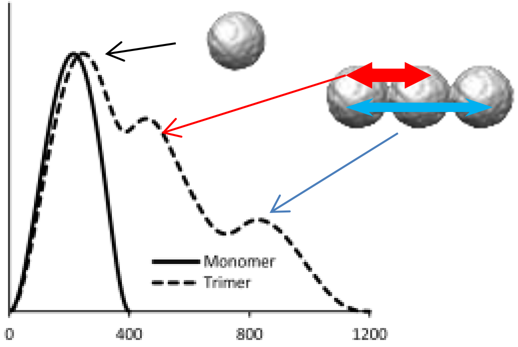
P(r) provides real space
Rg and Dmax.



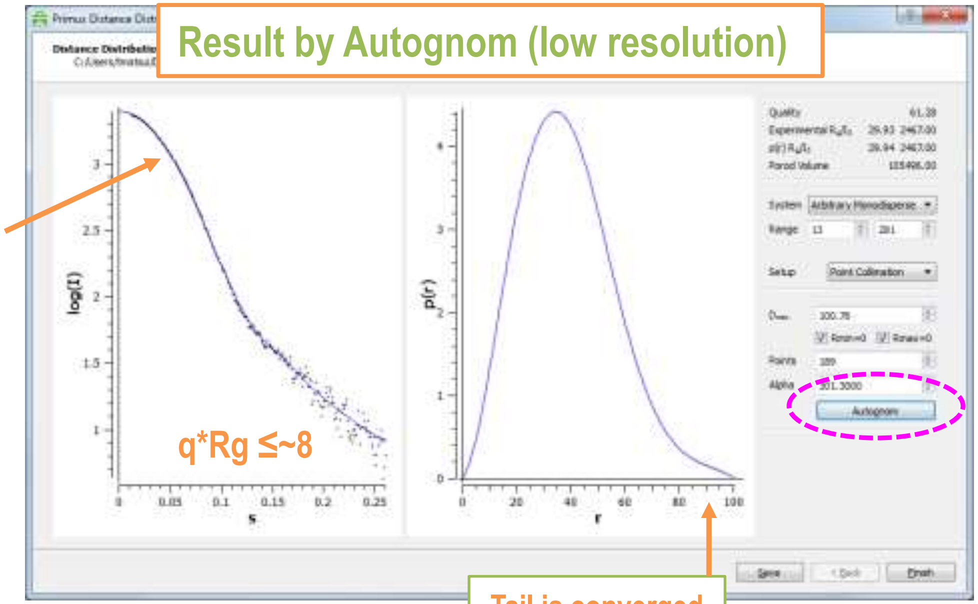
P(r) is the SAXS
analog of
crystallographic
Patterson function.



Putnam, et al. 2007 Quarterly Reviews of Biophysics 40, 191



P(r), example using primusqt (ATSAS)



Result by Autognom (low resolution)

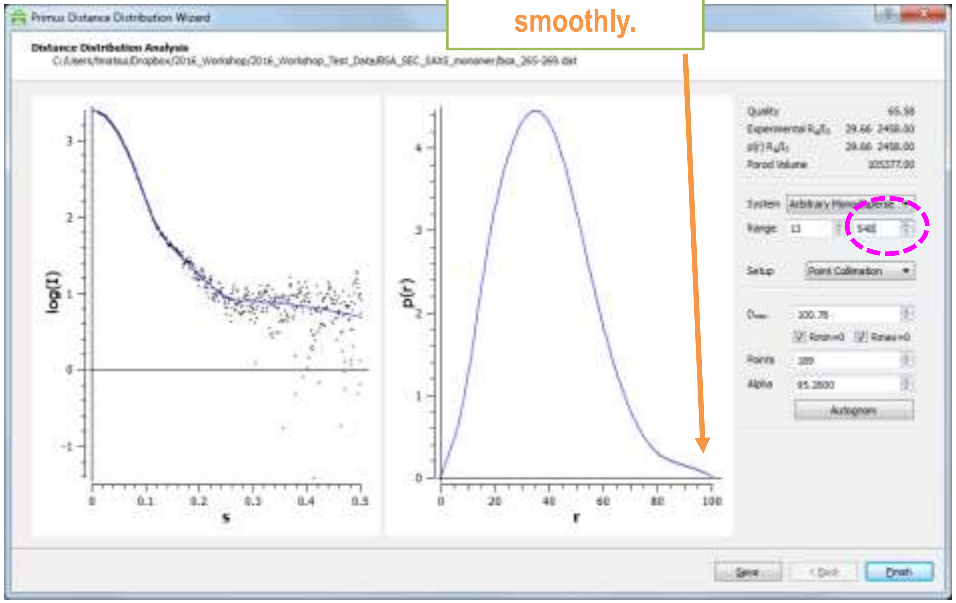
Perfect fitting from low resolution part.

q*Rg ≤~8

Quality 61.28
Experimental R_g/I_0 29.93 2467.00
p(r) R_g/I_0 29.94 2467.00
Porod Volume 103496.00
System: Arbitrary Monodisperse
Range: 13 281
Setup: Point Collimation
D_{max}: 100.78
☑ Rmin=0 ☑ Rmax=0
Points: 189
Alpha: 301.3000

Autognom button

Tail is converged smoothly.



Distance Distribution Analysis
C:/Users/hmatsui/Dropbox/2016_Workshop/2016_Workshop_Test_Data/BSA_SEC_SAXS_monomer/bsa_265-269.dat

Quality 65.98
Experimental R_g/I_0 29.66 2458.00
p(r) R_g/I_0 29.66 2458.00
Porod Volume 103377.00
System: Arbitrary Monodisperse
Range: 13 540
Setup: Point Collimation
D_{max}: 100.78
☑ Rmin=0 ☑ Rmax=0
Points: 189
Alpha: 85.2800

Expand q-range to high resolution.

Rg, Dmax and Porod volume are hardly influenced by resolution.

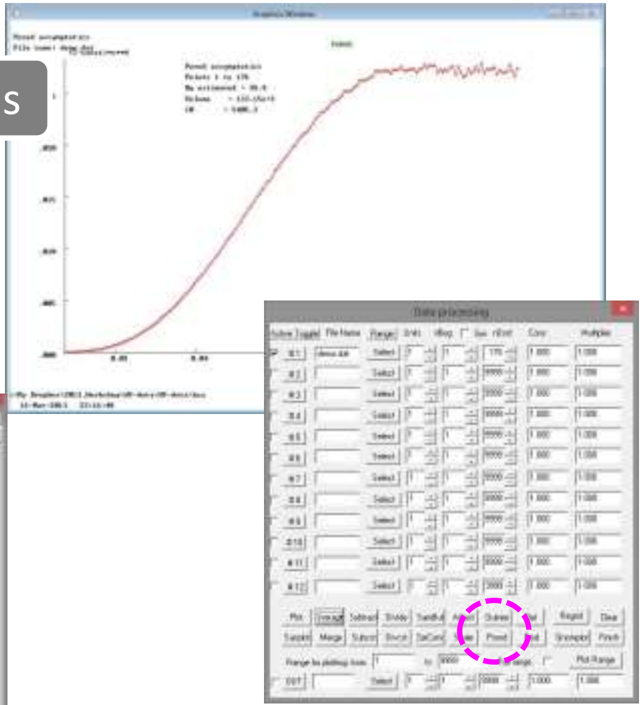
- More details will be provided in Q/A (hands-on) sessions.

Molecular weight (Mw) estimation

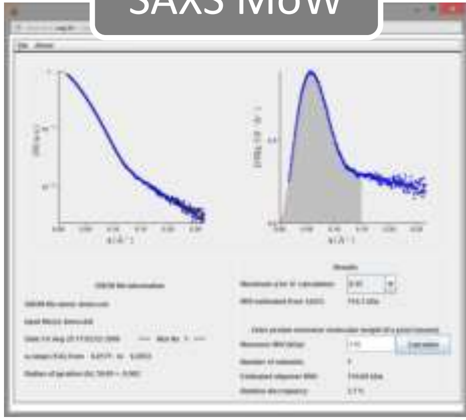
- **Standard curve**
  $Mw_{exp} / I(0)_{exp} \approx Mw_{standard}/I(0)_{standard}$
- **Water scattering (available in Graphit)**
  Orthaber *et al.*, J Appl Cryst (2000) 33, 218
- **SAXS MoW***
  JAVA applet (http://www.if.sc.usp.br/~saxs/)
- **Porod volume***
  Mw ≈ Porod volume *0.625
- **DATVC/DATMOW (ATSAS)***
- **Volume of dummy atoms***
  Mw ≈ (Dammif or Dammin volume)/2

***scale independent**

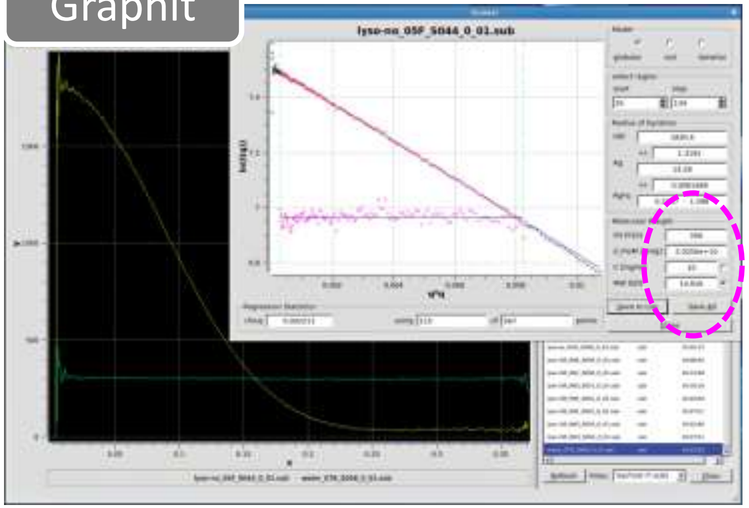
Primus

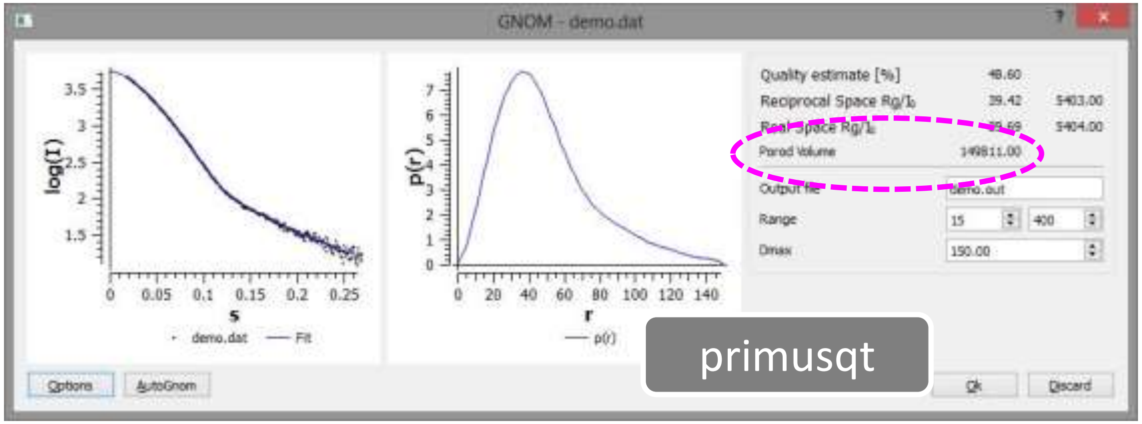


SAXS MoW



Graphit





Quality estimate [%] 48.60
Reciprocal Space Rg/I0 39.42 5403.00
Real Space Rg/I0 39.69 5404.00
Porod Volume 149811.00
Output file demo.out
Range 15 400
Dmax 150.00

primusqt

Acknowledgements

SSRL BL4-2

Thomas M. Weiss

Ivan Rajkovic

Ping Liu

All collaborators

All BL4-2 users

SLAC
NATIONAL ACCELERATOR LABORATORY

SSRL

Structural Molecular Biology

U.S. DEPARTMENT OF
ENERGY

SERVING SCIENCE AND SOCIETY • DOE Biological and Environmental Research Program

National Institute of
General Medical Sciences