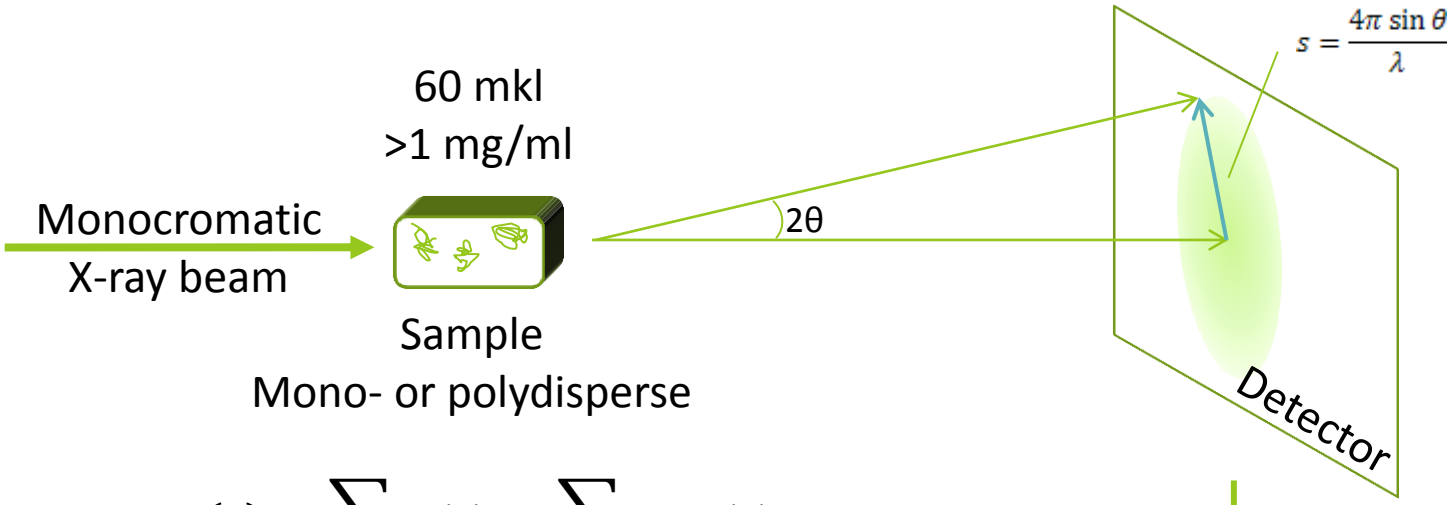
Development of Novel Small-Angle X-ray Scattering Data Analysis Methods for Study of Flexible Proteins

Michael Kachala

EMBL-Hamburg, Germany

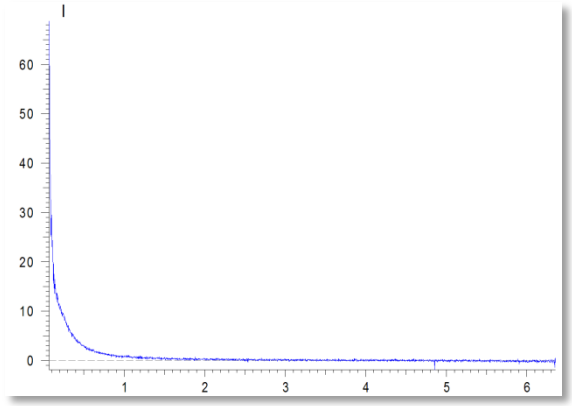
IDPbyNMR ITC, Les Houches
13 September 2012

Small Angle X-ray Scattering



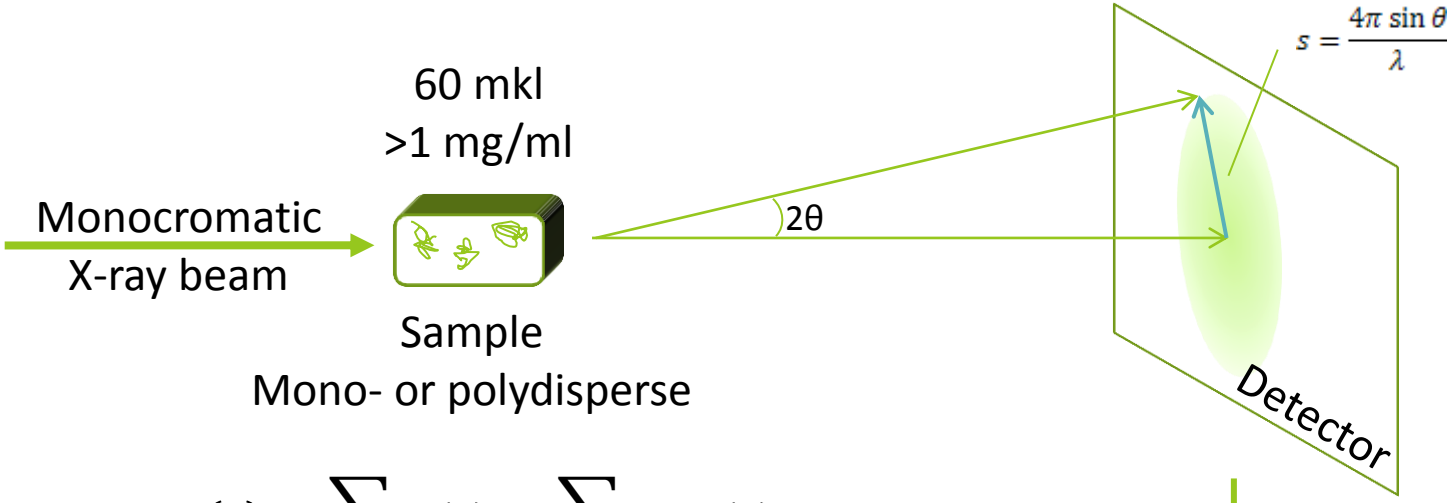
$$I(s) = \sum_N I_j(s) = \sum_K \nu_k I_k(s)$$

$$I(s) = \langle I(s) \rangle_\Omega = \langle A(s) \cdot A^*(s) \rangle_\Omega$$



Data analysis with ATSAS

Small Angle X-ray Scattering

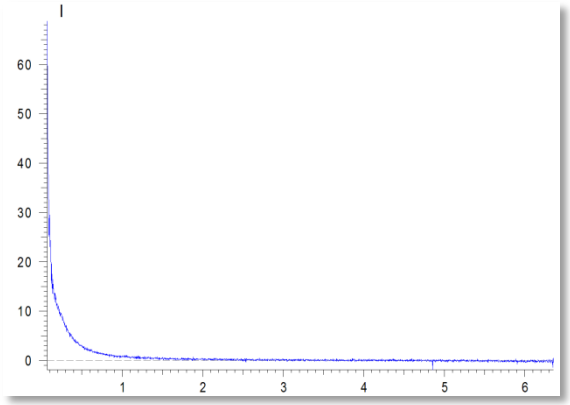


$$I(s) = \sum_N I_j(s) = \sum_K \nu_k I_k(s)$$

$$I(s) = \langle I(s) \rangle_\Omega = \langle A(s) \cdot A^*(s) \rangle_\Omega$$

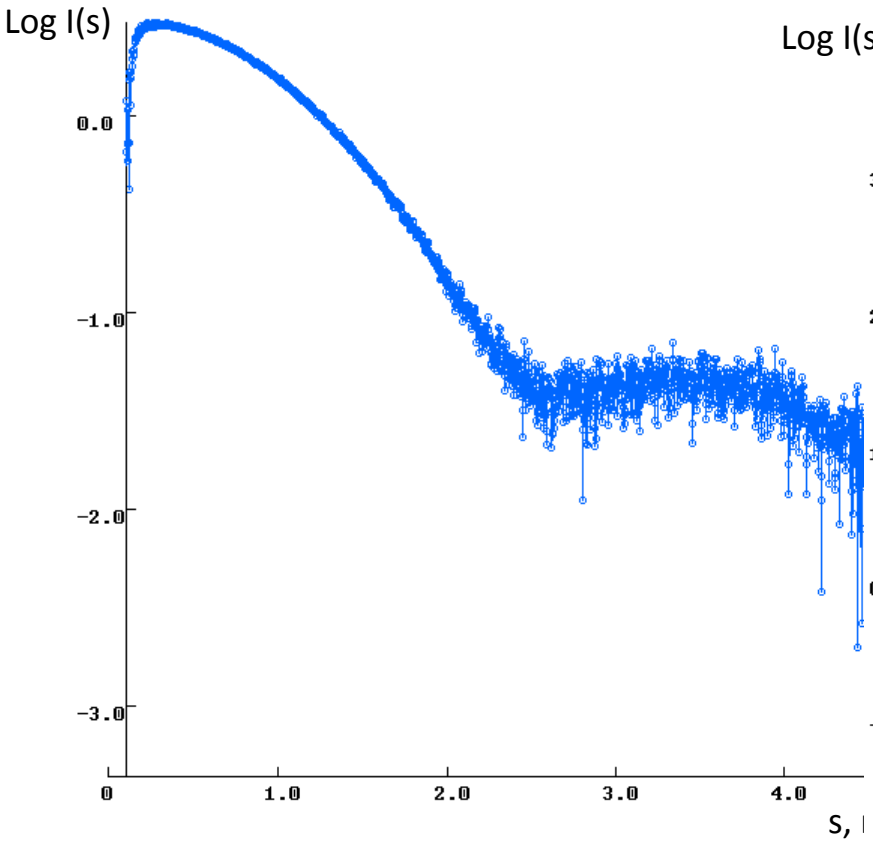
Methods to study IDPs:	
NMR	High resolution Limited size of the protein
SAXS	**Overall parameters** **Fast measurement – a lot of various conditions**
FRET	Single-molecule approach Protein must be modified

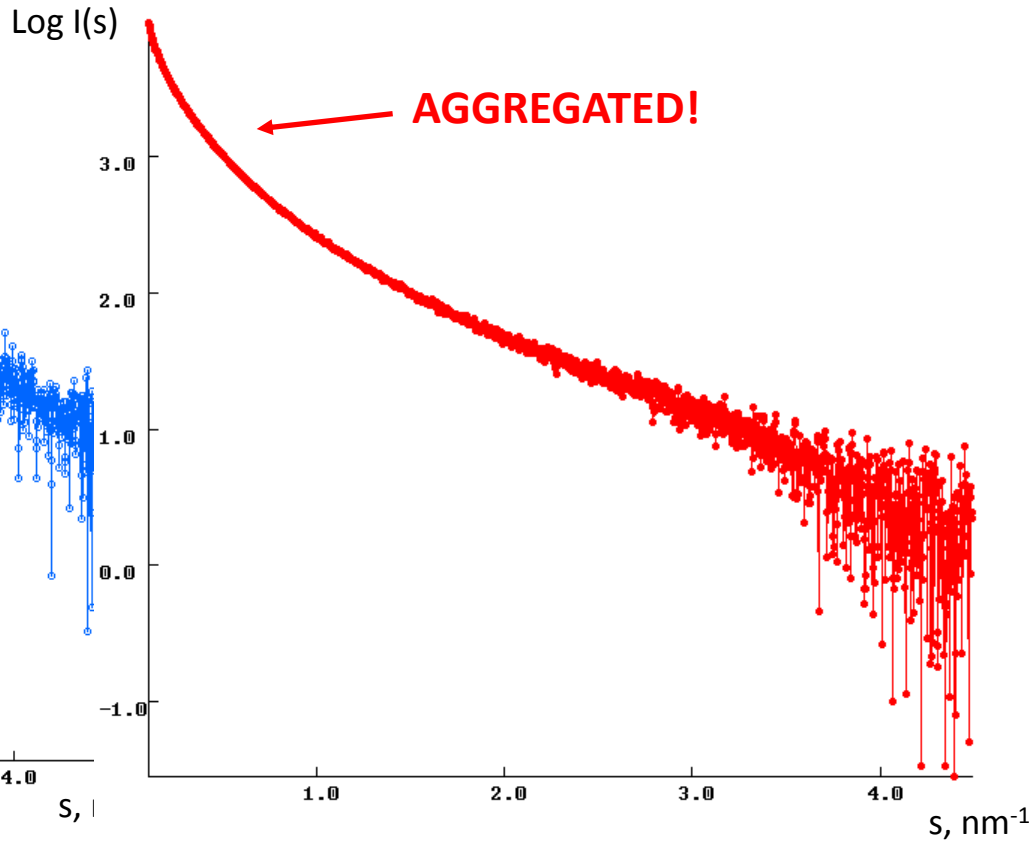
Data analysis with ATSAS



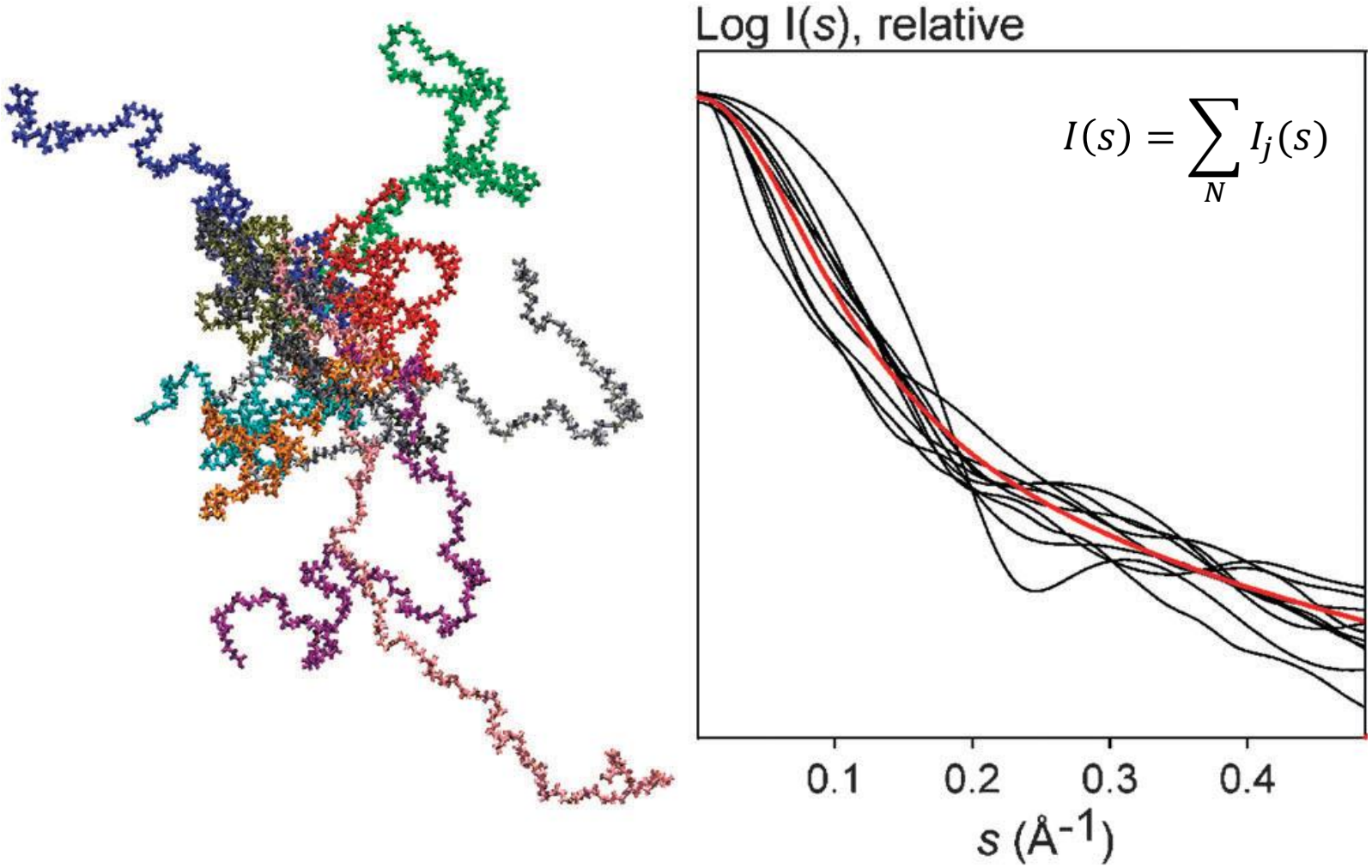
Initial Data Analysis: Data Quality

"Can I use this data for further analysis?"





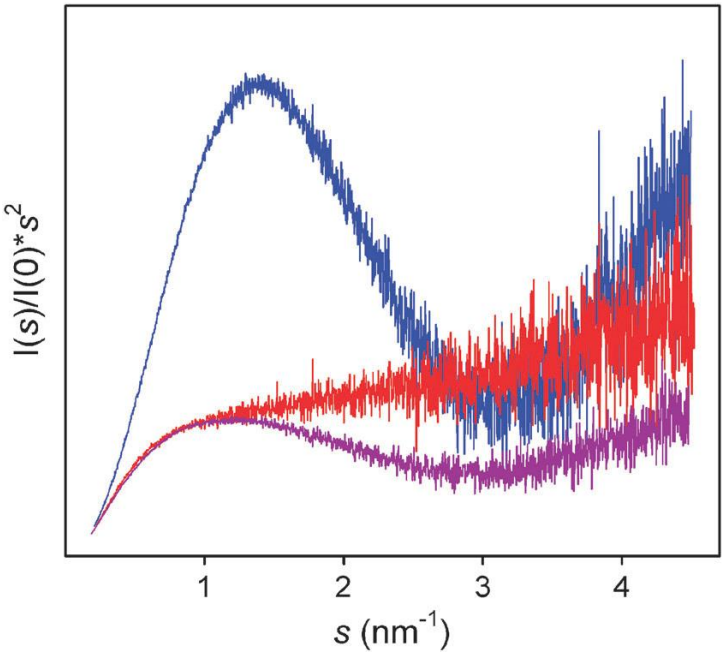
IDPs Scattering Curves



Source: Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering
Pau Bernadó and Dmitri I. Svergun Mol. BioSyst., 2012,8, 151-167

Using SAXS for IDPs structural characterization

1. Kratky plot analysis ($I(s) \cdot s^2$ vs s)



Kratky plot for three constructs of Src-Kinase

Folded

Unfolded

Both Folded and Unfolded

2. Radius of gyration is a single parameter
3. More comprehensive analysis – **Ensemble Optimization Method (EOM)**

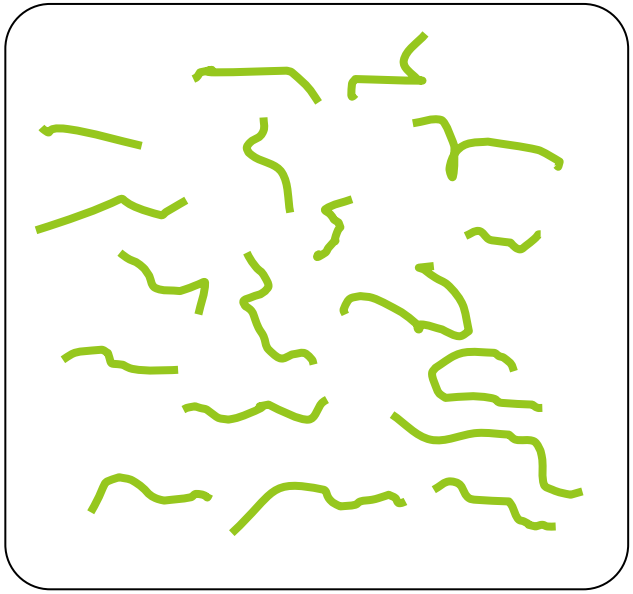
Source: Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering
Pau Bernadó and Dmitri I. Svergun Mol. BioSyst., 2012,8, 151-167

Ensemble Optimization Method (EOM)

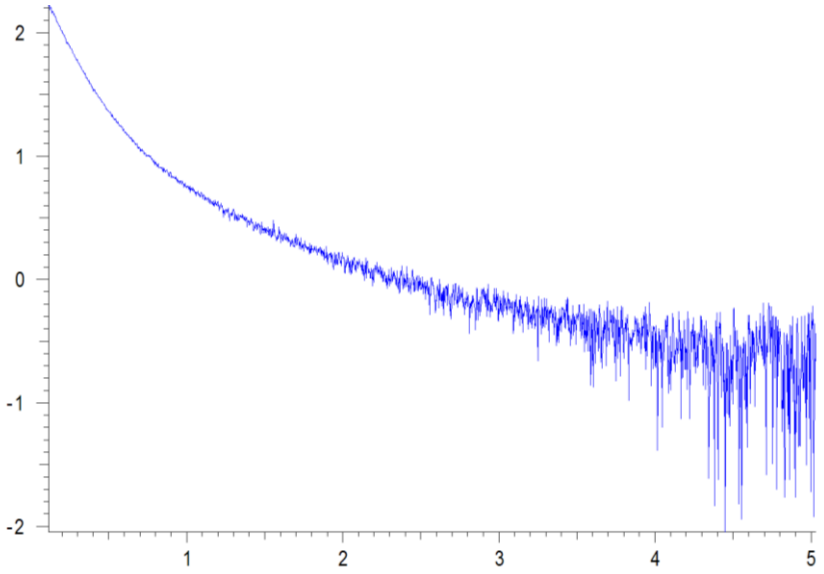
EOM represents sample as ensemble of structures which fits experimental data selected by genetic algorithm from randomly generated pool

Protein sequence,
Domains (if any),
Pool size, etc.

Experimental
scattering curve



Pool of randomly
generated structures



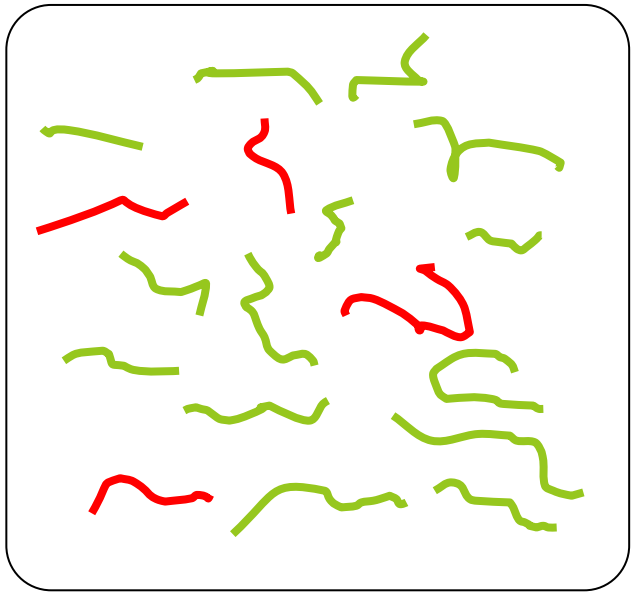
Fitting of experimental data
using genetic algorithm

Ensemble Optimization Method (EOM)

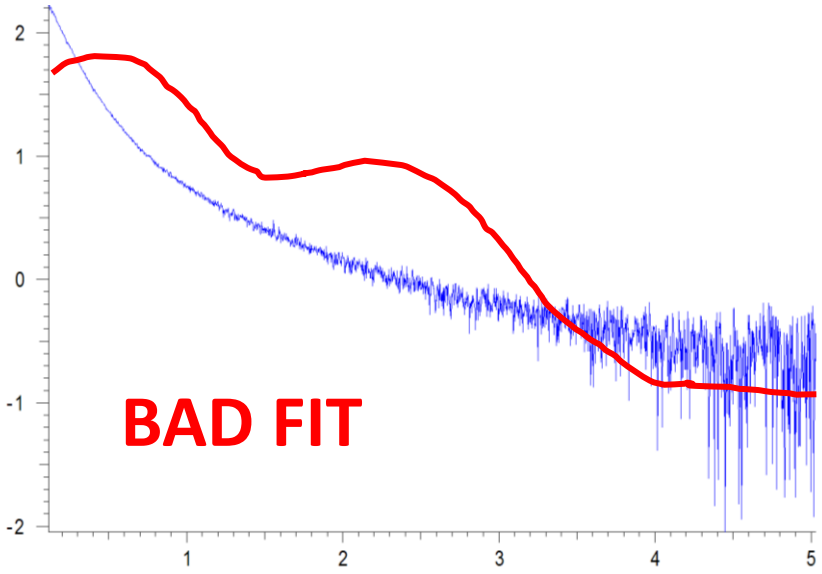
EOM represents sample as ensemble of structures which fits experimental data selected by genetic algorithm from randomly generated pool

Protein sequence,
Domains (if any),
Pool size, etc.

Experimental
scattering curve



Pool of randomly generated structures



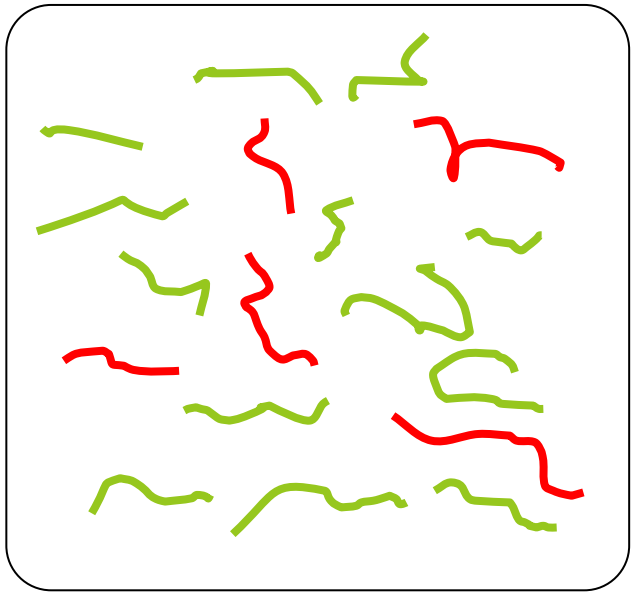
Fitting of experimental data using genetic algorithm

Ensemble Optimization Method (EOM)

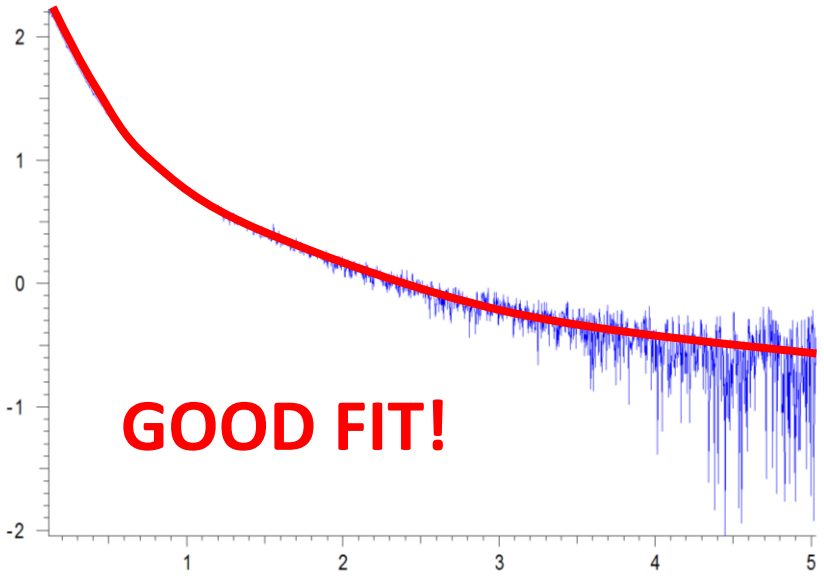
EOM represents sample as ensemble of structures which fits experimental data selected by genetic algorithm from randomly generated pool

Protein sequence,
Domains (if any),
Pool size, etc.

Experimental
scattering curve

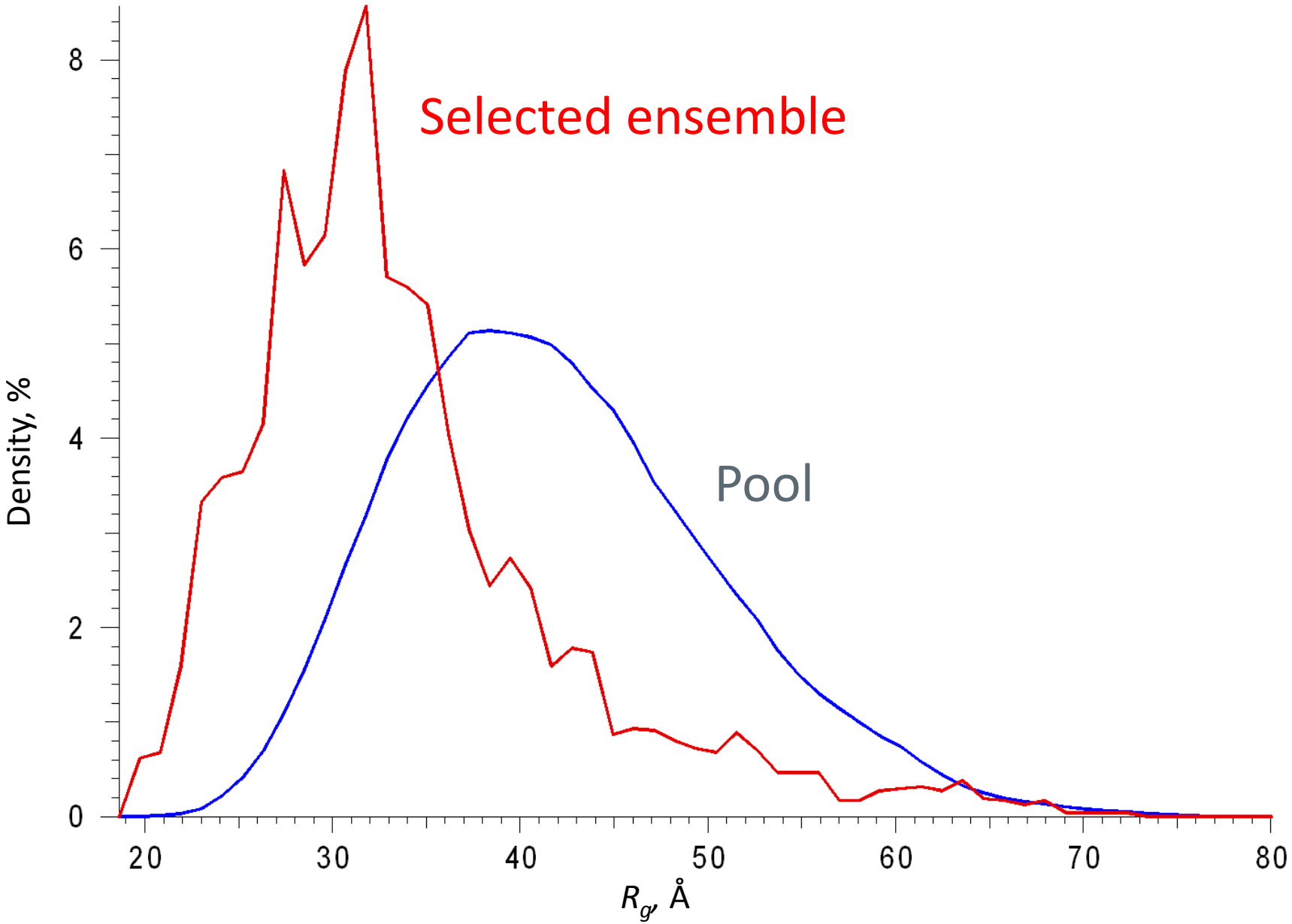


Pool of randomly generated structures



Fitting of experimental data using genetic algorithm

EOM: Rg Distribution



Examples of EOM application

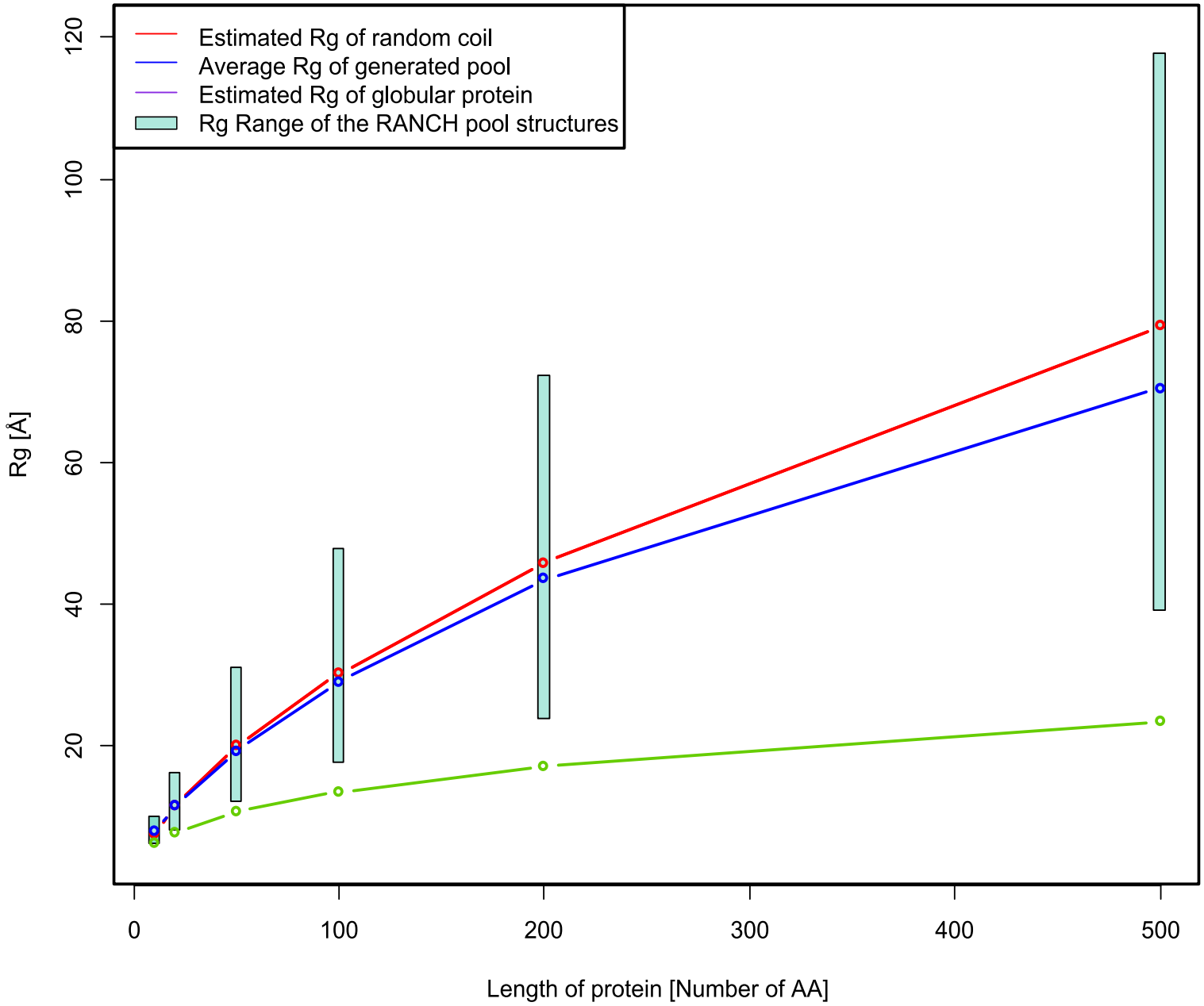
- Shkumatov, A. V., S. Chinnathambi, et al. (2011). "Structural memory of natively unfolded tau protein detected by small-angle X-ray scattering." Proteins 79(7): 2122-2131.
- Morgan, H. P., H. D. Mertens, et al. (2012). "Structural analysis of the C-terminal region (modules 18-20) of complement regulator factor H (FH)." PLoS One 7(2): e32187.
- Devarakonda, S., K. Gupta, et al. (2011). "Disorder-to-order transition underlies the structural basis for the assembly of a transcriptionally active PGC-1alpha/ERRgamma complex." Proc Natl Acad Sci U S A 108(46): 18678-18683.
- El Houry Mignan, S., G. Witte, et al. (2011). "Characterization of the chipsi subcomplex of Pseudomonas aeruginosa DNA polymerase III." BMC Mol Biol 12: 43.
- Kazantsev, A. V., R. P. Rambo, et al. (2011). "Solution structure of RNase P RNA." RNA 17(6): 1159-1171.
- Kim, H. S., M. C. Wilce, et al. (2011). "Different modes of interaction by TIAR and HuR with target RNA and DNA." Nucleic Acids Res 39(3): 1117-1130.

And many more…

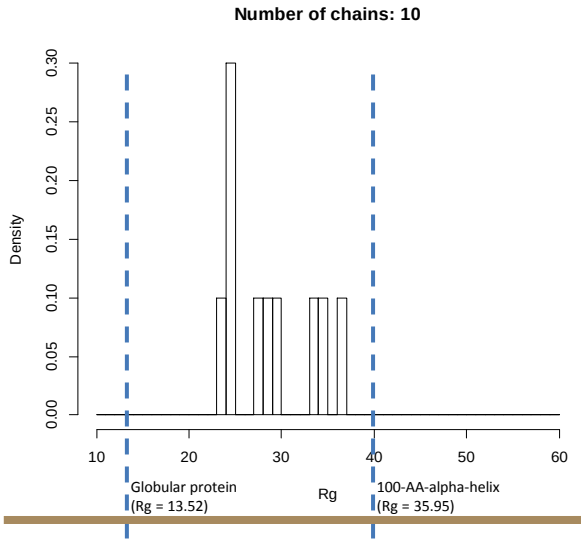
Questions to Answer

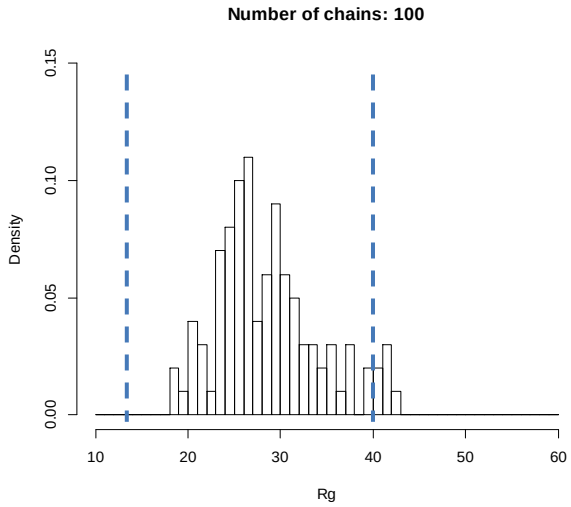
- What are the properties of generated structures?
- What are the optimal parameters for EOM?
- What are possible application of EOM?
- How results obtained using EOM can be used for quantitative characterization of disordered proteins?

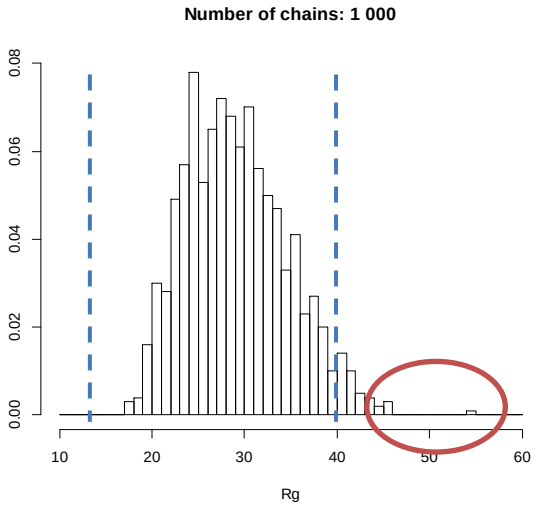
EOM Tests: Rg of the Generated Structures

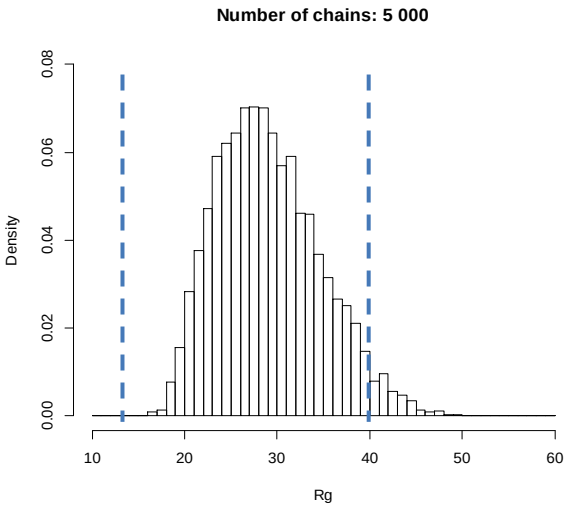


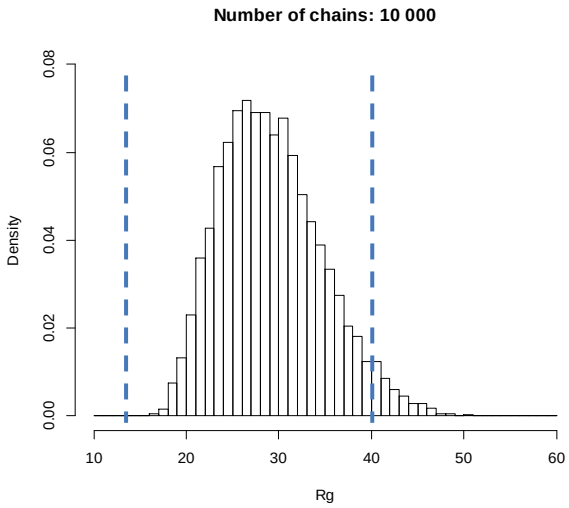
EOM Tests: Size of Pool

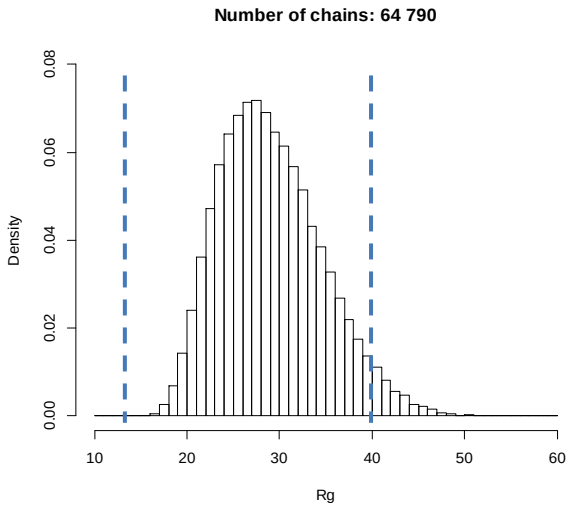




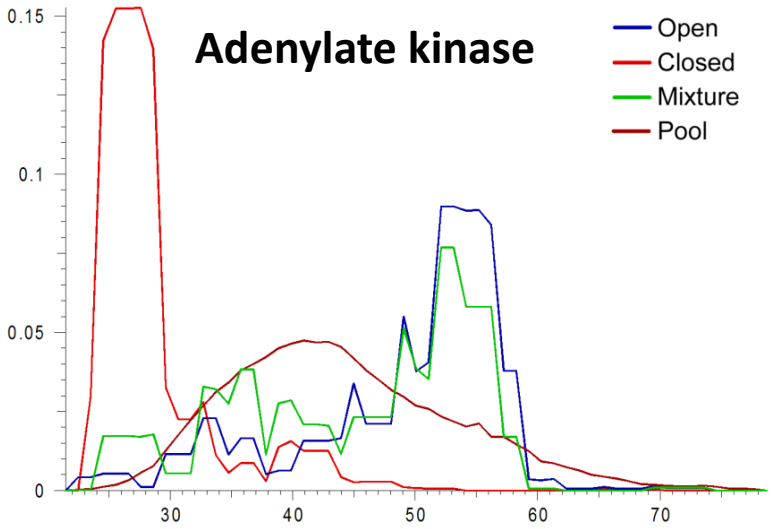


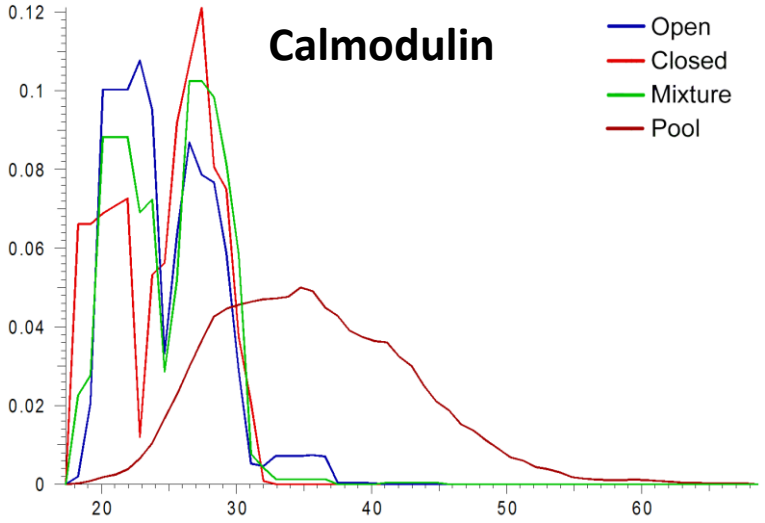


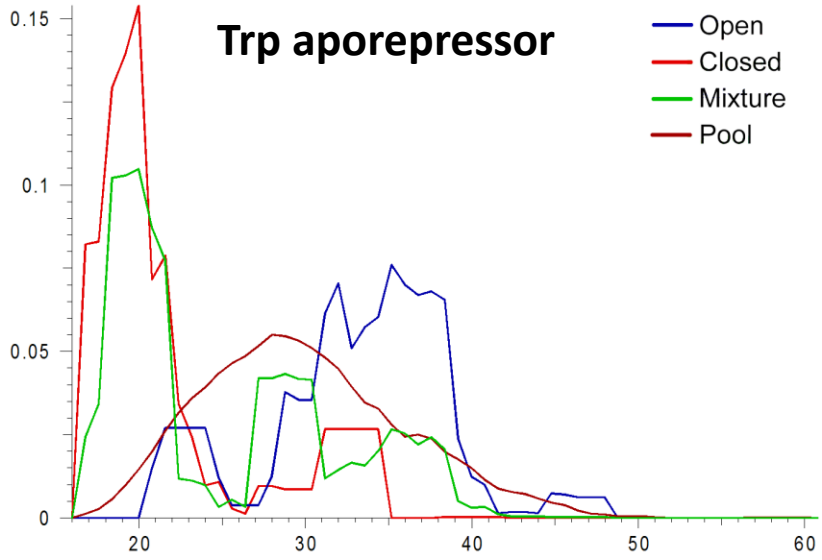




EOM Tests: Resolving Open and Closed Conformations





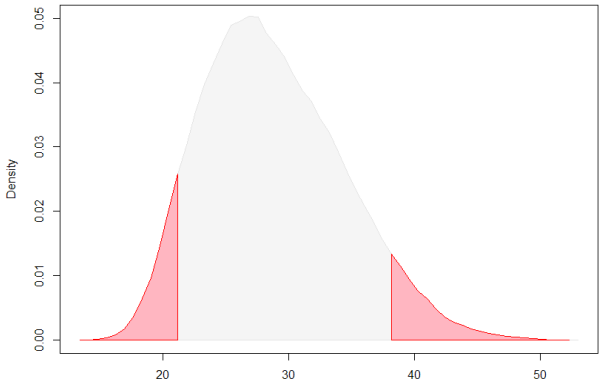


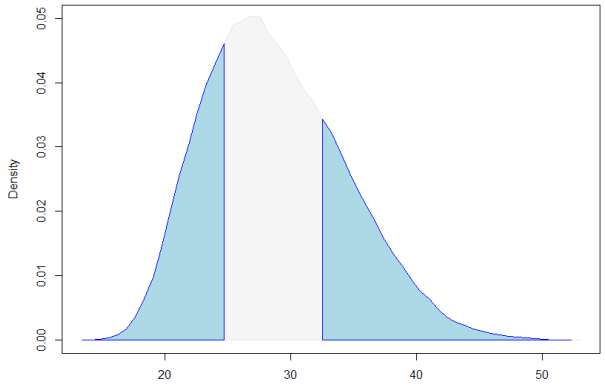
Proteins	R_g, open conformation (CRYSOL)	R_g, closed conformation (CRYSOL)	Min R_g of the pool	Max R_g of the pool
Adenylate kinase	20.88	17.98	24.82	75.92
Calmodulin	21.53	16.29	20.84	66.56
Tryptophan repressor	30.61	18.59	18.40	58.38

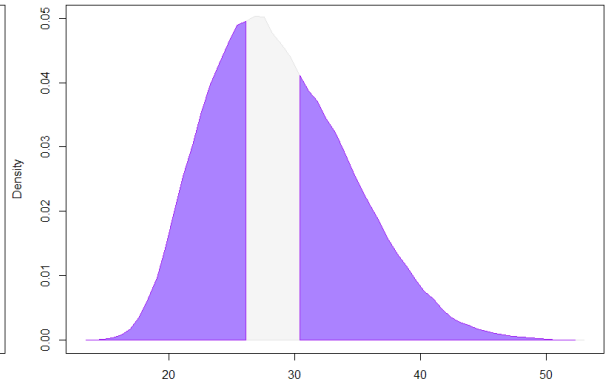
Resolution of Subpopulations by EOM

- Generate a pool, select two subpopulations from the it and calculate scattering curve for their union

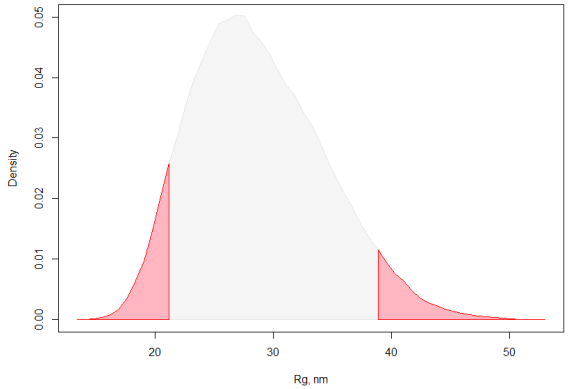
Wide subpopulations

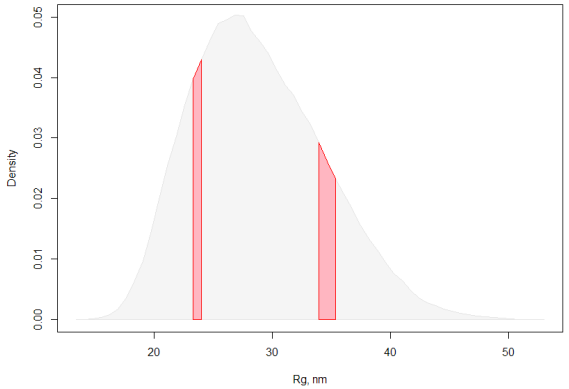


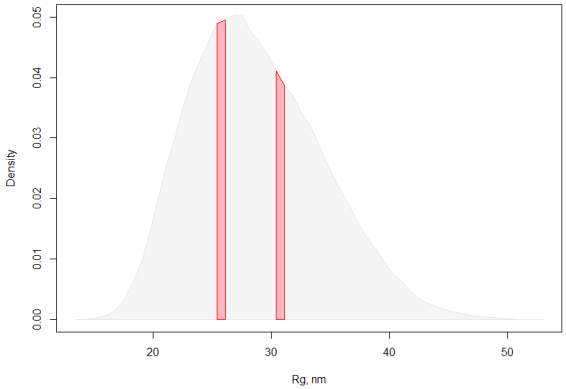




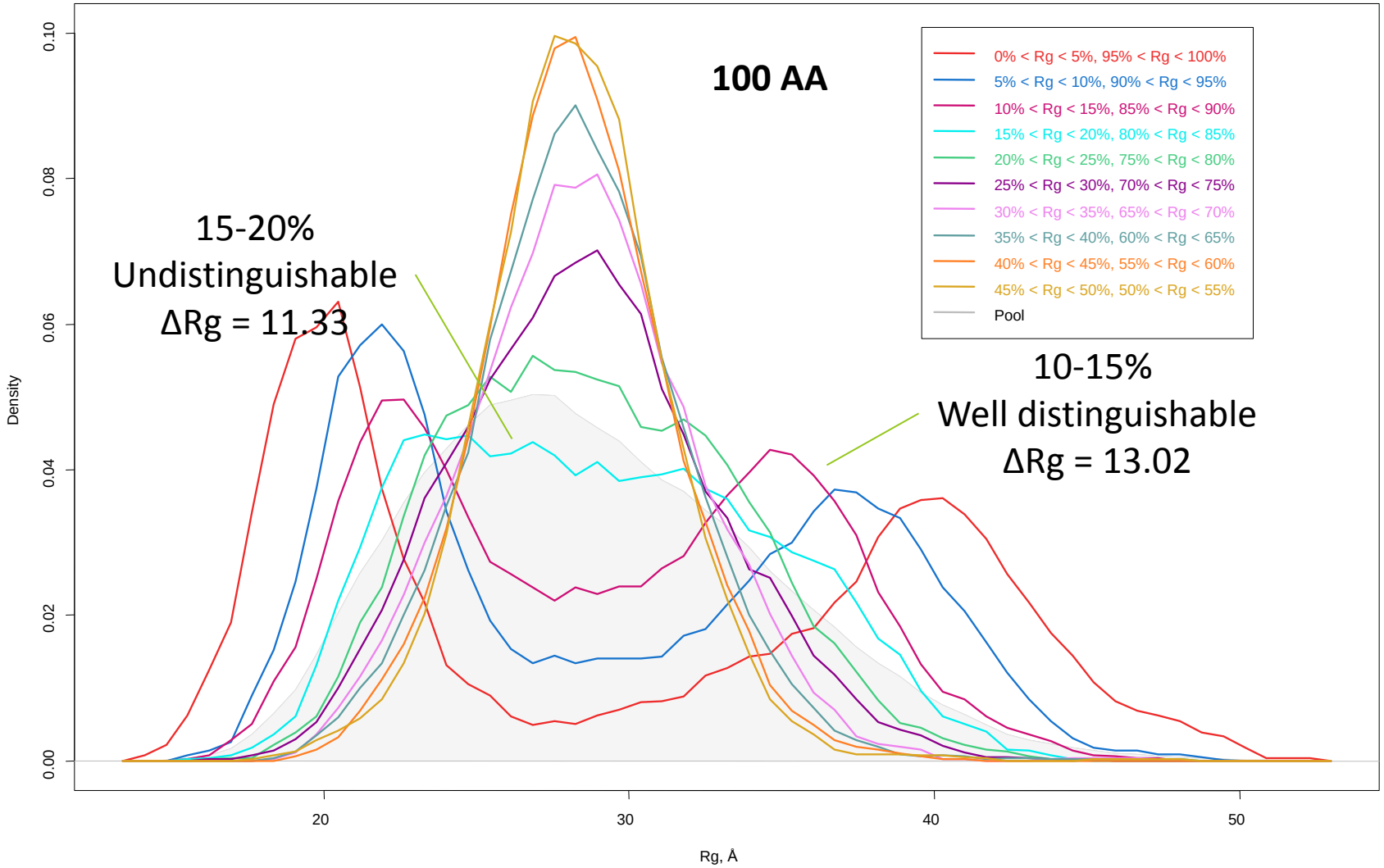
Narrow subpopulations





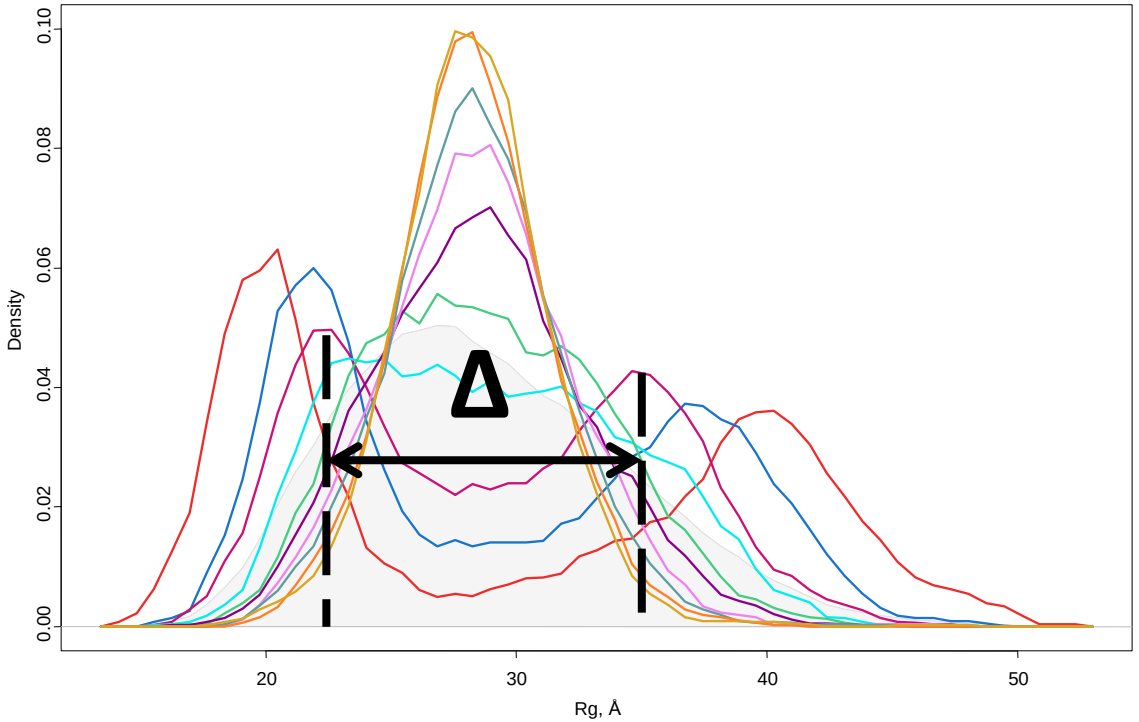


Resolution of Subpopulations by EOM



Concept of EOM Resolution

- **Absolute Resolution** (Δ) is the minimal difference in average R_g between two populations still distinguishable on Rg distribution

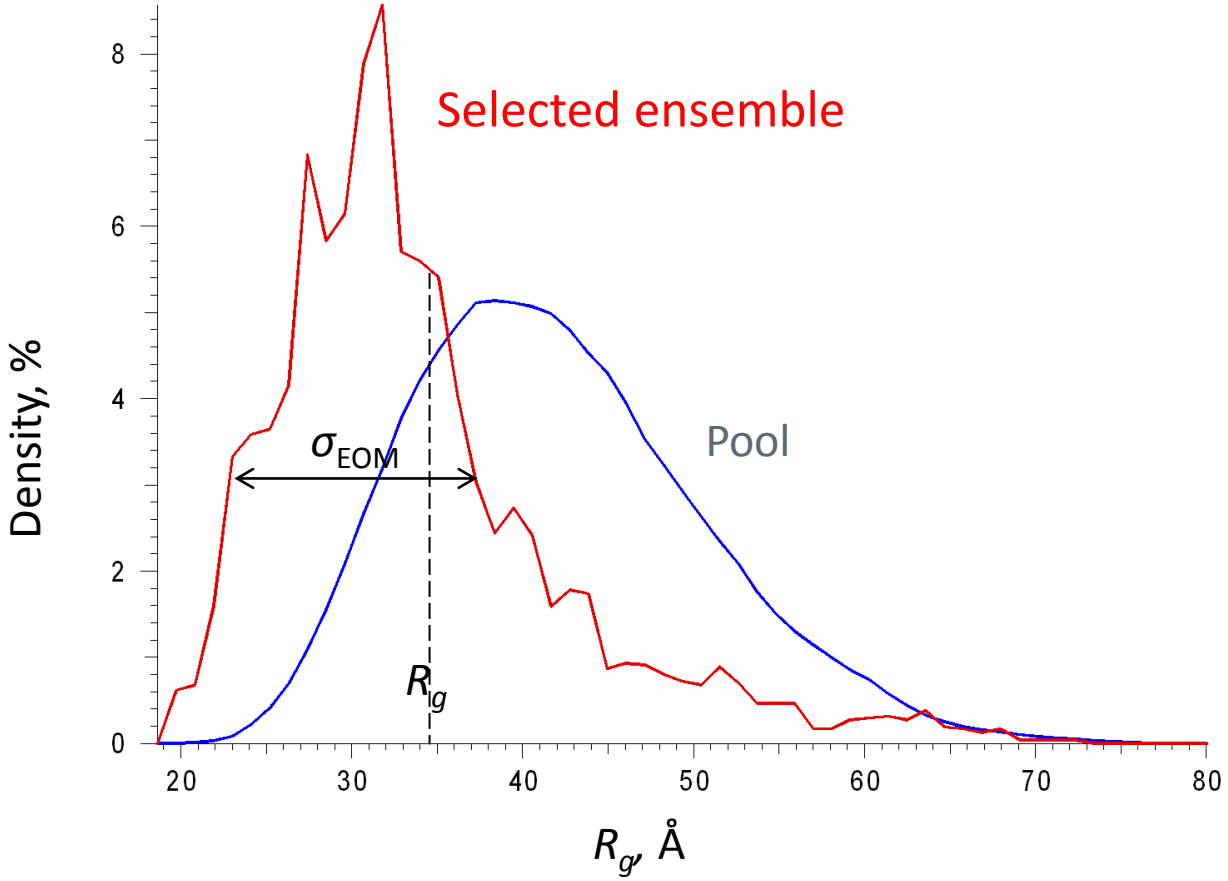


- **Relative Resolution (δ)** is ratio between Resolution and standard deviation of the pool:

$$\delta = \Delta/\sigma$$

If **δ > 2.3** then EOM can be used to distinguish two populations

Quantitative Parameters



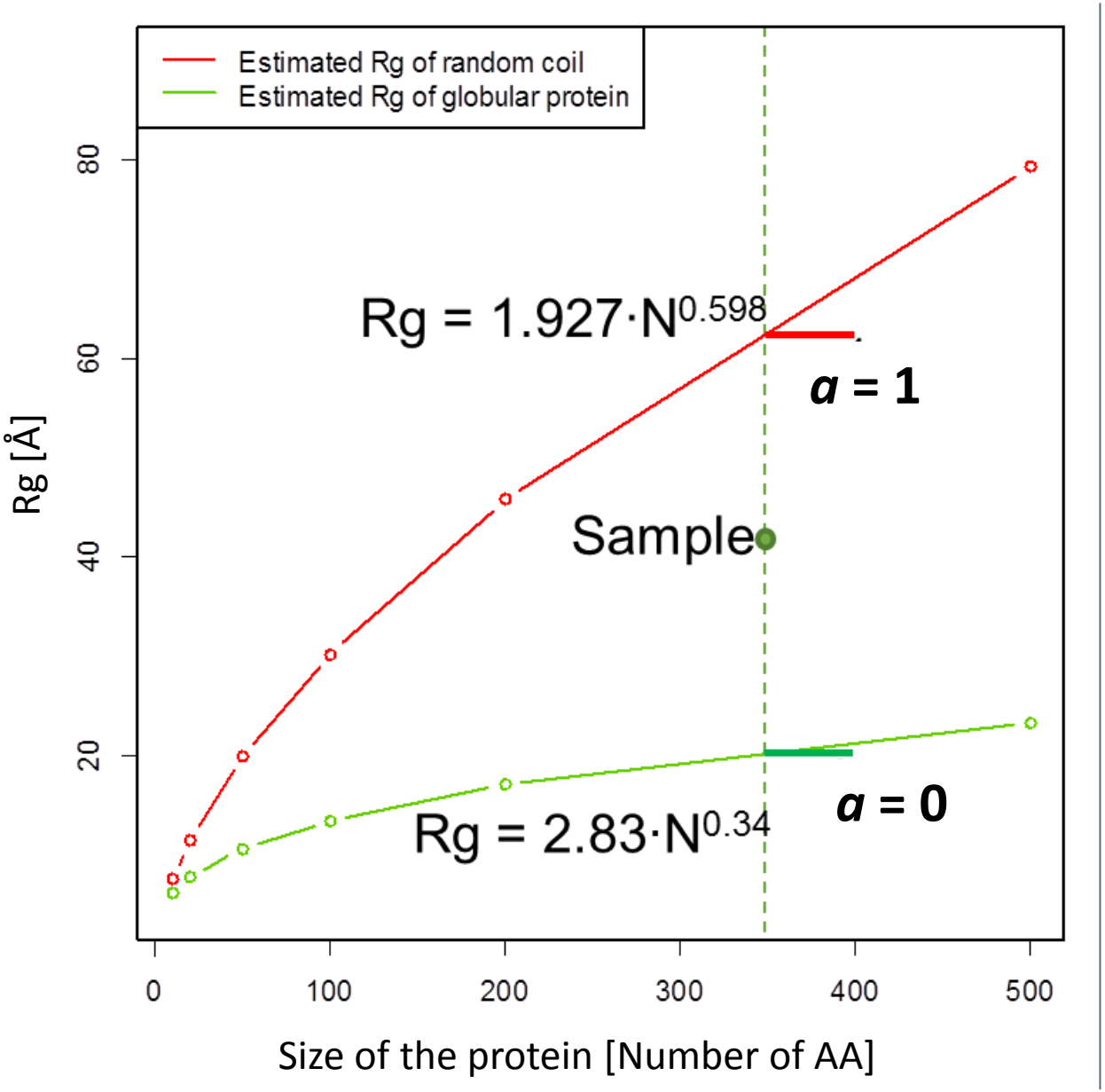
Absolute values

- Radius of gyration (Rg)
- Width of GAJOE distribution (σ)

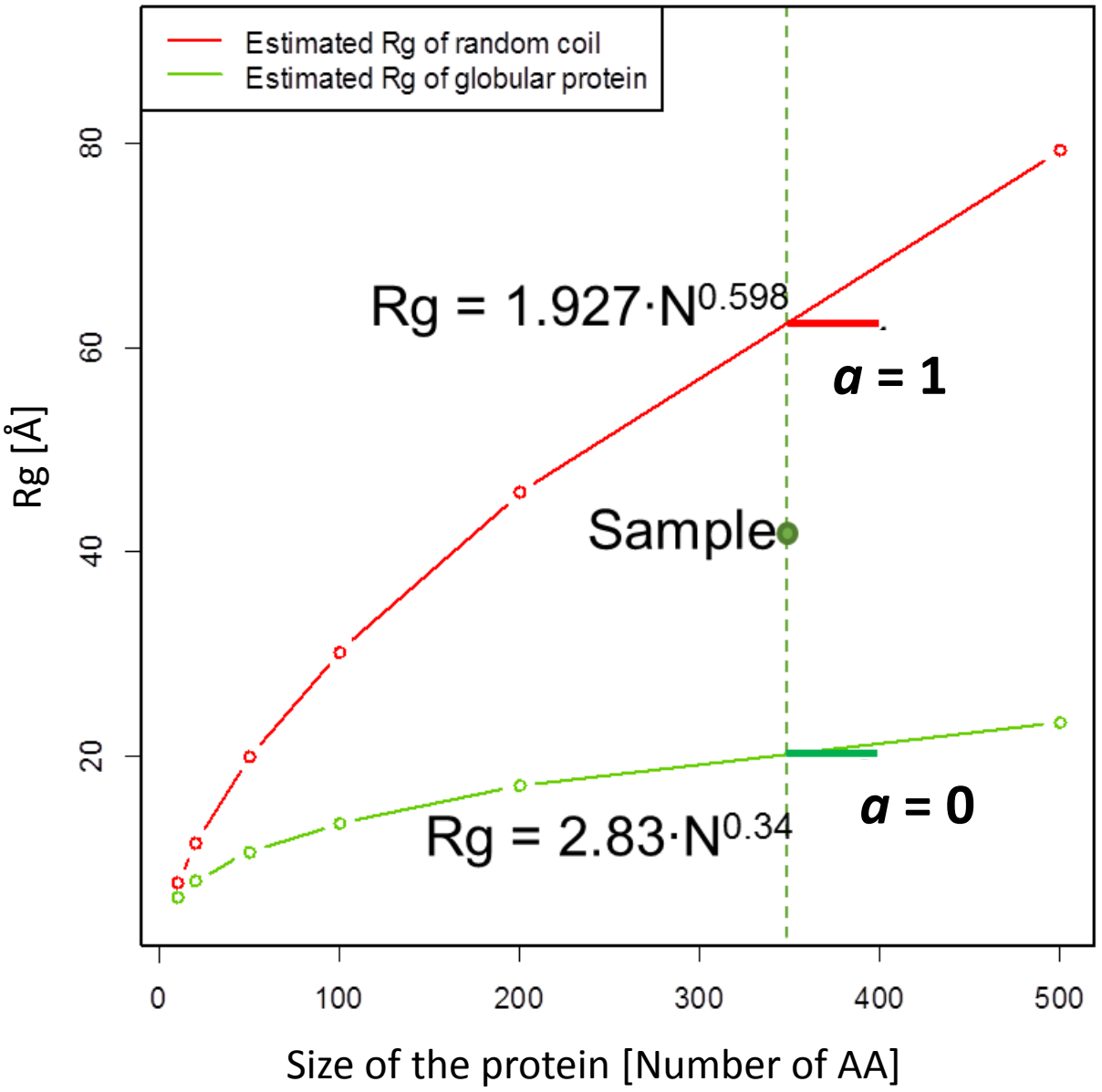
Relative parameters

- $a = f(R_g, N)$
- $b = \sigma_{EOM}/\sigma_{pool}$

Quantitative parameters: Determination of *a* and *b*

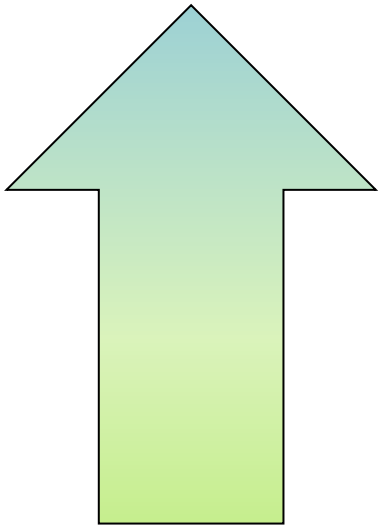


Quantitative parameters: Determination of *a* and *b*



b is determined by running EOM on synthetic datasets for rigid structures

Flexible as the pool ***b* = 1**

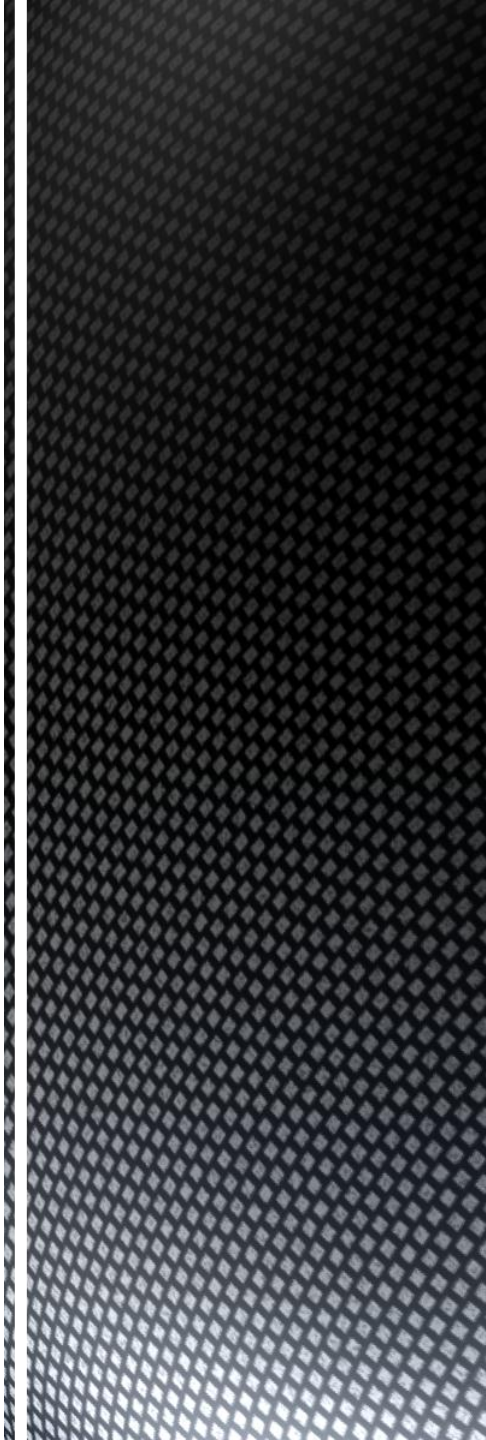


Rigidity border is ***b* = 0.64**

Conclusions

- EOM can be used as tool of SAXS data analysis for both flexible and rigid structures, determination of level of flexibility, distinguishing between two populations or conformations
- EOM has some limitations that must be taken into account such as limited ability to generate pool structures with R_g close to globular protein and certain width of selected ensemble distribution for single structure
- Quantitative parameters to estimate samples flexibility are proposed based on the R_g and standard deviation of R_g for selected ensemble

Thank you for your attention!



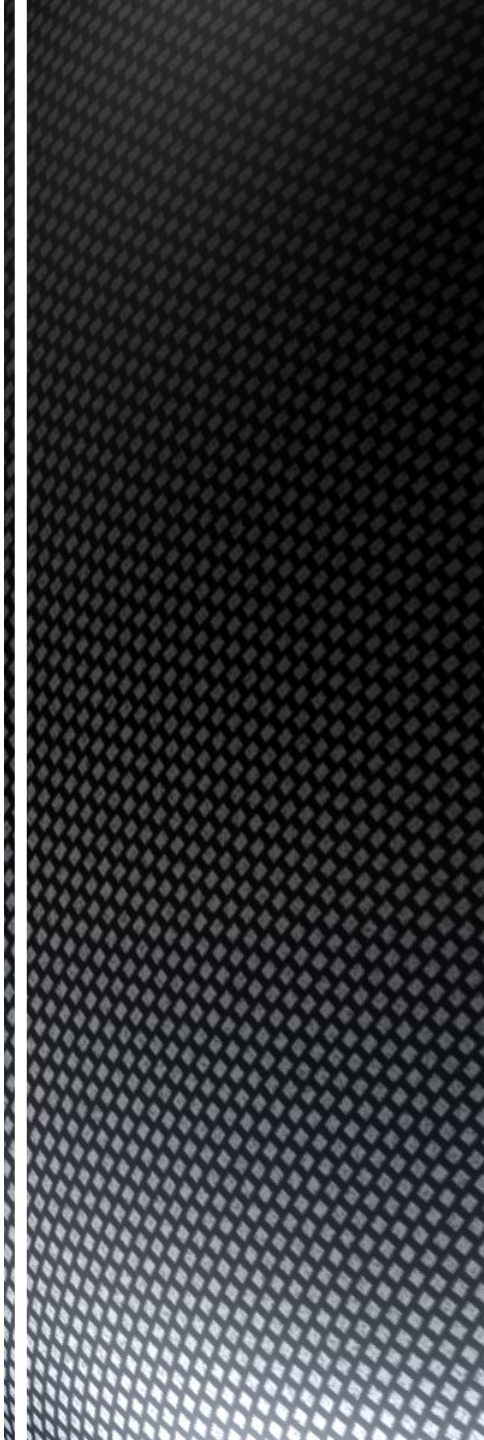
…And don't forget about

IDPbyNMR ITC on SAXS and Computational Techniques

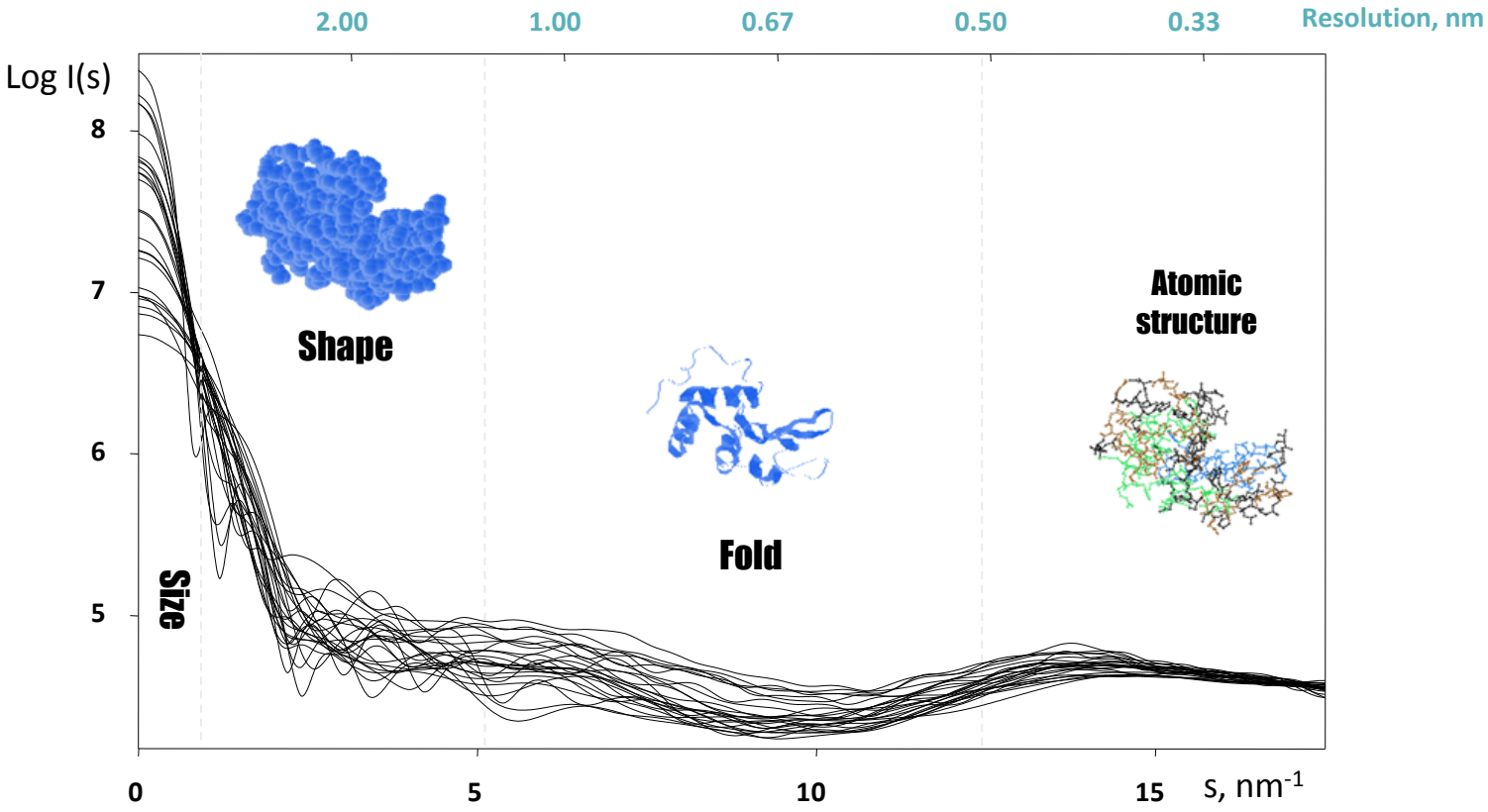
EMBL Hamburg, Germany

Monday 11 March - Saturday 16 March 2013

Backup slides

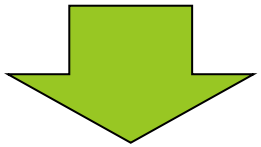


Initial Data Analysis: Data Range



Concept of EOM Resolution

1. Capability to distinguish two subpopulation does not depend on their width
2. This capability depends on the difference between average R_g of subpopulations
3. Minimal difference depends on the length of the polypeptide chain chosen for pool generation



- **Absolute Resolution** (Δ) is the minimal difference in average R_g between two populations still distinguishable on GAJOE Rg distribution
- **Relative Resolution** (δ) is ratio between Resolution and standard deviation of the pool:

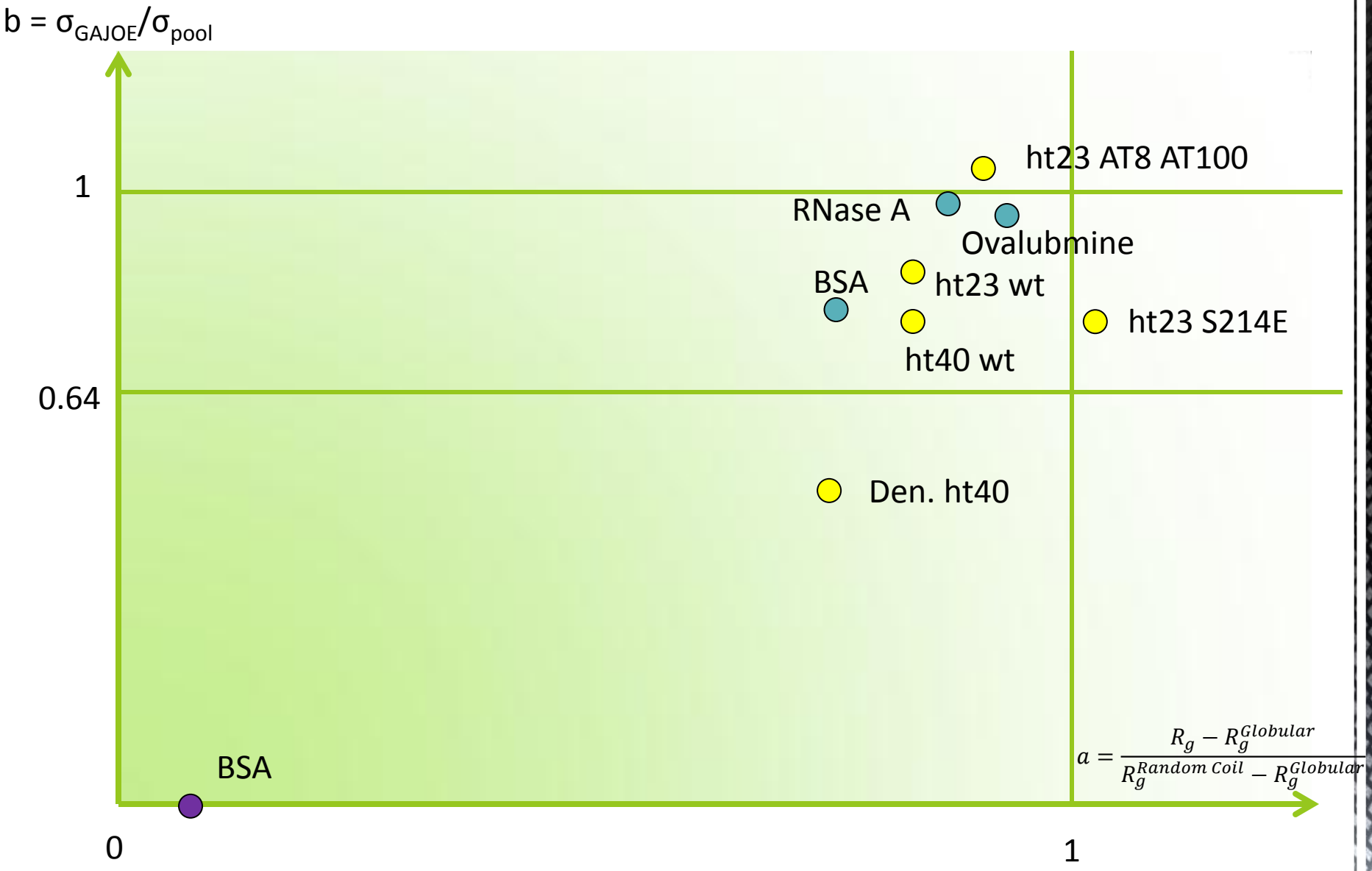
$$\delta = \Delta/\sigma$$

Estimation of the resolution:

Case	σ, Å	Smallest distinguishable ΔR_g < Δ < Largest undistinguishable ΔR_g, Å	δ estimation
100AA, wide	5.52	10.86 < Δ < 12.85	1.97 < δ < 2.32
100AA, narrow	5.52	11.33 < Δ < 13.02	2.05 < δ < 2.36
200AA, narrow	8.52	16.21 < Δ < 19.90	1.90 < δ < 2.34

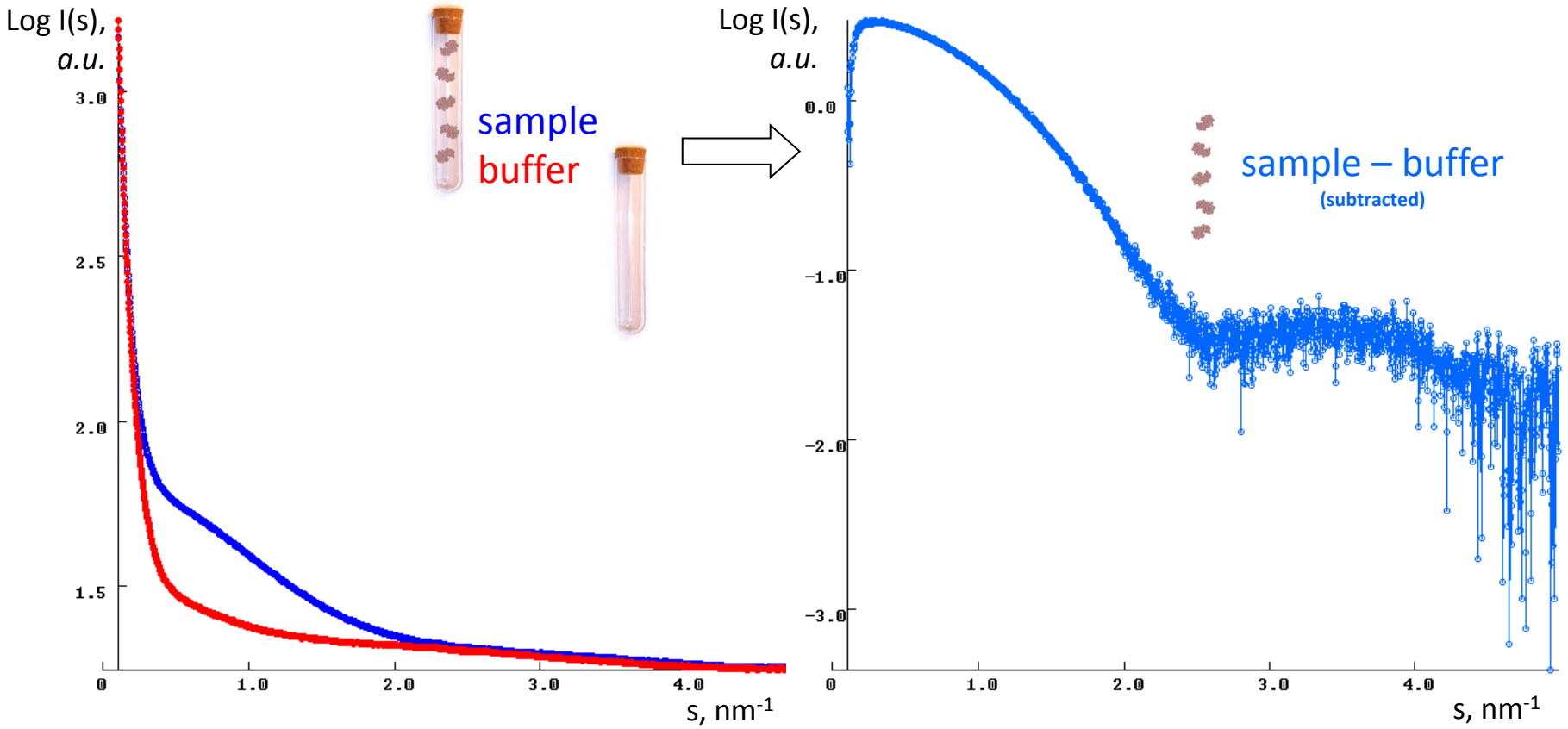
If δ is more than **2.3** then GAJOE can be used to distinguish two populations

Quantitative Parameters: Real-life Cases



Initial Data Analysis: Background Subtraction

Solution* minus *Solvent



Looking for protein signals **less than 5%** above background level

Genetic Algorithm