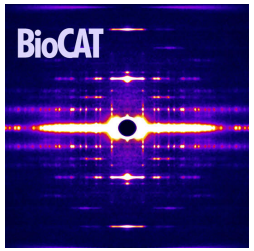


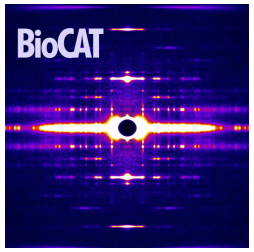
Analyzing flexible and disordered systems using SAXS

Jesse Hopkins, PhD
IIT/CSRRI
Staff Scientist, BioCAT
Sector 18, Advanced Photon Source

ILLINOIS INSTITUTE OF TECHNOLOGY

October 29, 2020

CSRRI



SAXS and flexibility/disorder

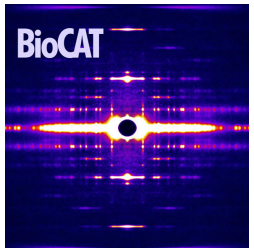
- Flexible systems adopt a continuous range of conformations
  - Distinct from polydisperse systems that adopt a small number of distinct states (conformations, oligomers)

- Measured SAXS profile is a combination of all the different species/conformations in solution
  - Volume weighted sum of all the individual scattering profiles

$$I(q) = \sum_{n=1}^{K} \nu_n I_n(q)$$

- For flexible systems, this is an advantage, as it means a single SAXS measurement samples the entire conformational ensemble

- SAXS is one of the few methods that can quantitatively characterize partially disordered or completely disordered macromolecules
  - Often used in combination with NMR

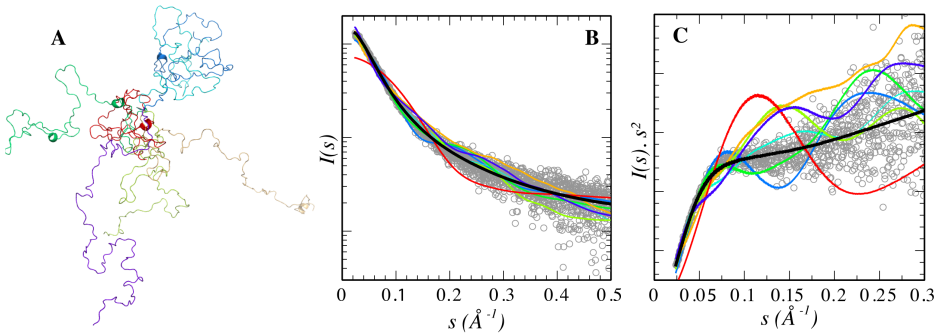


Characteristics of flexibility in SAXS

- How can you tell if you're measuring a flexible system?
  - Use multiple metrics
- Characteristics of a flexible system's scattering profile:
  - Smooth I(q), with little or no fine structure
  - Plateau or increase in dimensionless Kratky plot
  - No plateau in Porod-Debye plot (completely disordered)
  - Porod exponent < 4
- Characteristics of the P(r) distribution:
  - Smooth P(r) with little or no fine structure
  - Extended tail on P(r) function
  - D_{max} can be hard to determine
  - R_g, I(0) from P(r) usually larger than from Guinier
- Overestimates of M.W. weight from volumetric methods (Porod volume, envelopes)
  - Alternatively, low calculated density from Porod volume and known M.W.
- Other:
  - Guinier range may be narrow, $q_{max}R_g \sim 0.8$
  - Flory exponent close to ~0.6
  - Extended reconstructions or rigid body models

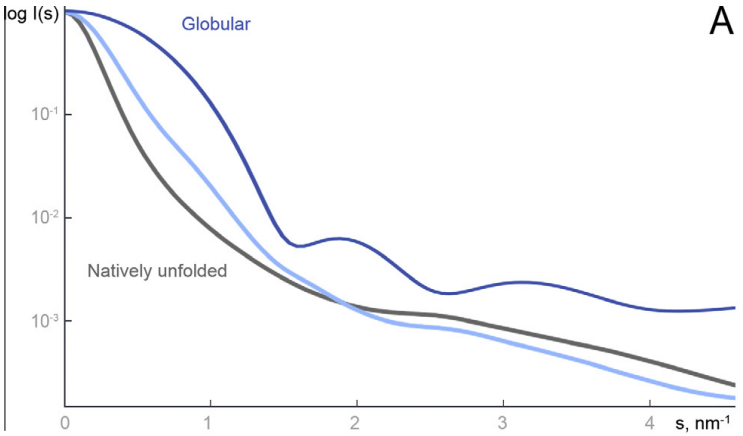
I(q) for flexible systems

- For fully disordered systems, I(q) is characteristically smooth, as it represent an average of a large number of conformations, which washes out distinct features



7 simulated conformers of an IDP, their scattering profiles, and the average scattering profile of 5000 conformers as compared to the data.

Cordeiro et al., 2017. DOI: 10.1007/978-981-10-6038-0_7



Simulated scattering profiles for a globular, 50% unfolded, and completely unfolded protein.

Kikhney and Svergun, 2015. DOI: 10.1016/j.febslet.2015.08.027

Porod exponent for flexible systems

- Porod's law (also Porod-Debye law) states that at high q, scattering intensity decays as:

$$I(q) \propto q^{-D}$$

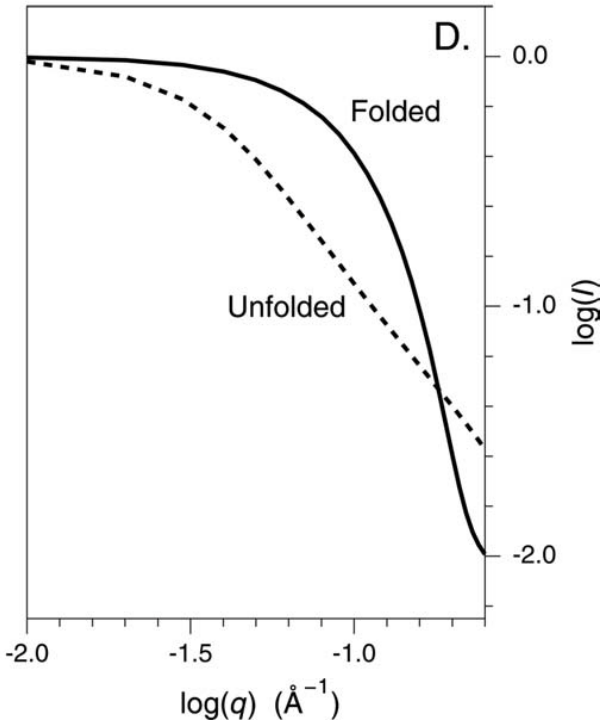
where D is the Porod exponent and depends on the particle shape/flexibility

- Porod exponents:
  - Hard sphere: 4
  - Rod: 3
  - Thin disc: 2
  - Random walk/gaussian chain: 2
  - Extended unfolded protein (self avoiding random walk/swollen gaussian chain): 1.7
  - Needle (fully extended chain): 1

- Smaller exponents indicate less globular systems, either flexibility or anisotropy

- Porod's law breaks down at higher q due to
  - Shape effects (folded/partly folded systems)
  - Hydration and excluded volume effects (all systems, $q \gtrsim 0.15$ Å^{-1})

- Fitting in the mid-q range, such as with ScAtter, can determine Porod exponent

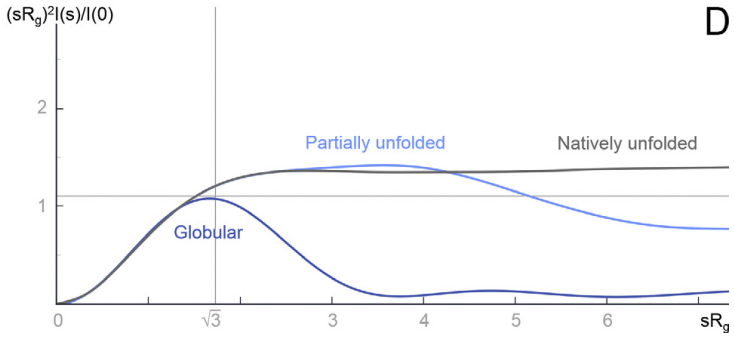


On a log-log plot, power laws like Porod's law look linear. You can see the folded system has a steeper slope at high q, translating to a larger Porod exponent.

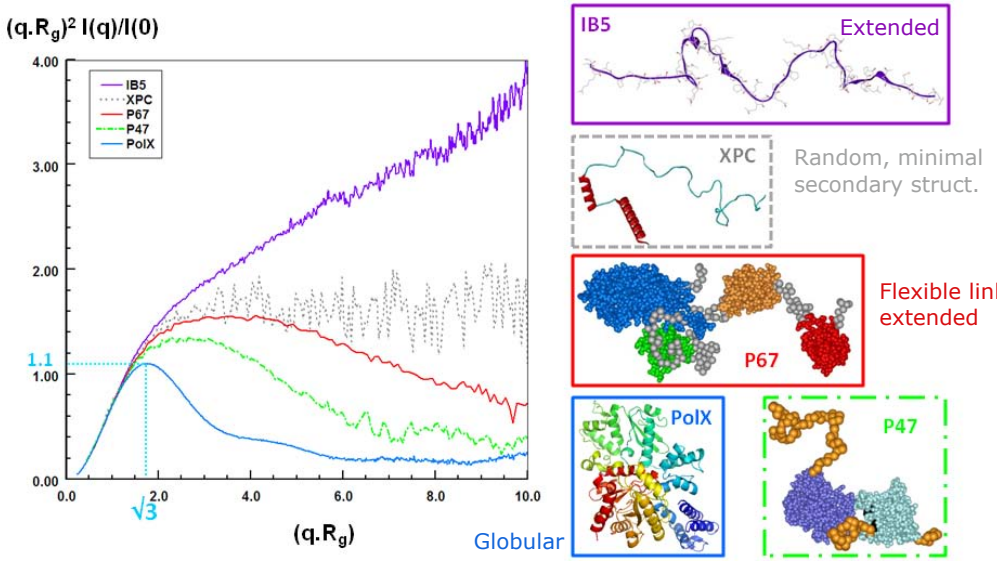
Johansen et al, 2011.
DOI: 10.1002/pro.739

Dimensionless Kratky plot

- Dimensionless Kratky plot: $(qR_g)^2I(q)/I(0)$ vs. qR_g
  - Removes effects of size, concentration to allow direct comparison of shape/conformation
- Globular systems have a characteristic peak of 1.104 at ~1.73 ($\sqrt{3}$)
- Random chains plateau between 1.5-2, may increase further at q > 0.2-0.3 $Å^{-1}$
- Fully extended chains increase beyond 2.0



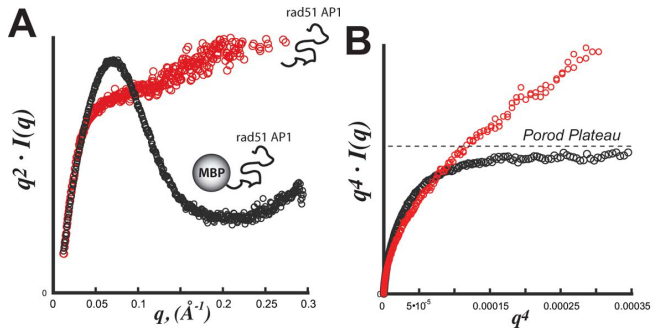
Kikhney and Svergun, 2015.
DOI: 10.1016/j.febslet.2015.08.027



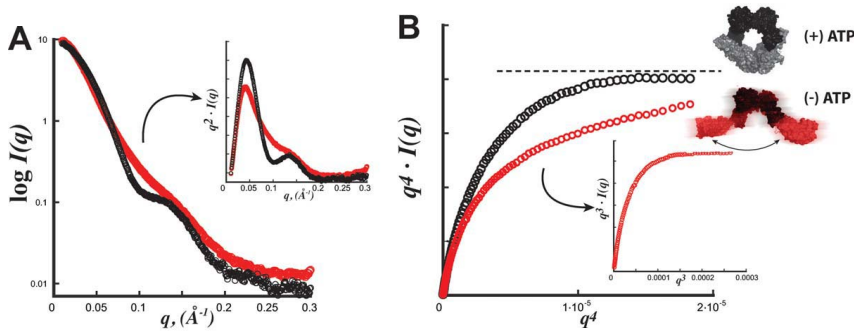
Receveur-Brechot and Durand, 2012.
DOI: 10.2174/138920312799277901

Porod-Debye plot

- Porod-Debye plot: $q^4I(q)$ vs. q^4
  - Look at low-to-mid q region, just after first peak of Kratky plot
- Can help distinguish between globular and flexible systems when Kratky plot is indeterminant
  - Use of $q^3I(q)$ vs. q^3 (sometimes 'SIBYLS' plot) can also help
- A plateau in the plot indicates a compact globular domain with minimal conformational flexibility in the system
  - Lack of a plateau can indicate full flexibility, or an extended object



IDP with and without globular MBP tag shows differences in unfolded vs. partly folded protein Kratky and Porod-Debye plots

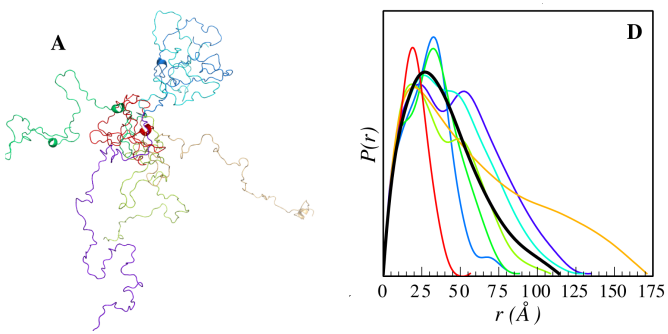


Kratky plot looks relatively globular for both ATP bound and unbound. However, Porod-Debye plot shows lack of plateau with ATP unbound. Together indicates well folded but flexible domains

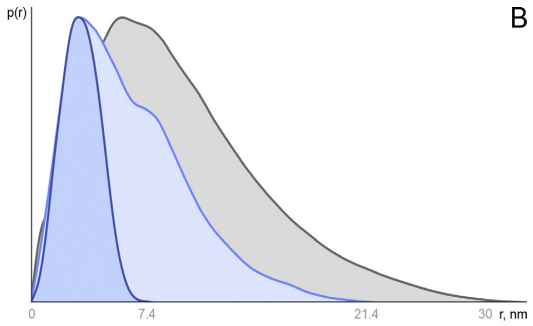
Rambo and Tainer, 2011. DOI: 10.1002/bip.21638

P(r) for flexible systems

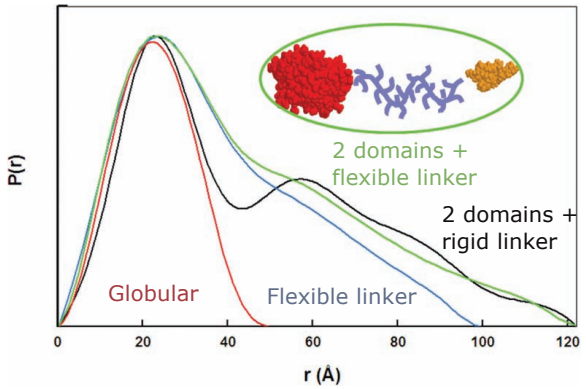
- For fully disordered systems, P(r) is characteristically smooth, as it represent an average of a large number of possible conformations, which washes out distinct features
- Extended tail on P(r) function gives slow approach to D_{max}
- D_{max} can be hard to determine
- R_g, I(0) from P(r) usually larger than from Guinier



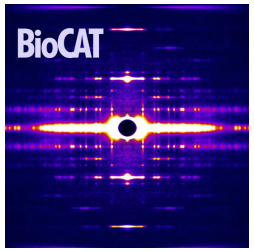
7 simulated conformers of an IDP, their P(r) functions, and the average P(r) of 5000 conformers. Cordeiro et al., 2017. DOI: 10.1007/978-981-10-6038-0_7



P(r) functions for simulated globular, 50% unfolded, and natively unfolded proteins. Kikhney and Svergun, 2015. DOI: 10.1016/j.febslet.2015.08.027



P(r) functions for a globular domain, a flexible linker, two domains plus a flexible linker, and a mutant with a rigid linker. Receveur-Brechot and Durand, 2012. DOI: 10.2174/138920312799277901



M.W. and density for flexible systems

- Flexible systems in solution occupy a larger volume, as measured by SAXS
- M.W. calculation that depends on volumes will tend to overestimate M.W. for flexible systems
  - Porod volume methods
  - M.W. estimated from 3D reconstructions
- If M.W. and oligomeric state of sample are known, can calculate observed particle density from Porod volume
  - Proteins: ~1.36 g/cm^3
  - Globular proteins via SAXS: ~1.3-1.4 g/cm^3
  - Flexible proteins via SAXS: ~0.9-1.3 g/cm^3
- M.W. differences for the volumetric methods vs. other methods, or a low calculated density in solution can indicate flexibility

Rambo and Tainer, 2011. DOI: 10.1002/bip.21638

Other indicators of flexibility

- Guinier fit may only extend to $q_{max}R_g$~0.8 (IDRs/IDPs)

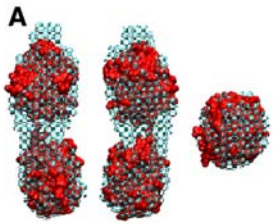
- Flory exponent ν

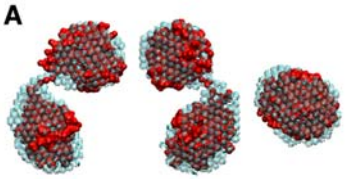
$$R_g = R_0 N^{\nu}$$

  - Defines relation between molecular size (R_g, end-to-end distance, average intra-residue spacing) and number of residues
  - $\nu = 1/D$ (Porod exponent)
  - Can be fit, assuming specific models for the system
  - IDPs: ~0.5-0.6
  - Globular: 0.33

- When doing reconstructions or rigid body modeling assuming 1 shape, results are characteristically extended, reconstructions show large volumes for flexible regions
  - High NSD not necessarily sign of flexibility, nor do flexible systems necessarily have high NSD
  - E.g. Reconstructing a two domain protein with flexible vs. fixed linker, flexible linker position not clearly visible

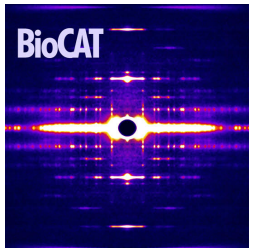
Flexible linker





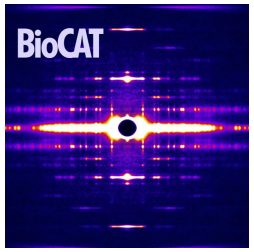
Rigid linker

Bernado, 2010. DOI: 10.1007/s00249-009-0549-3



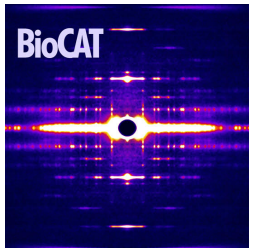
So is my system flexible?

- Indicators of flexibility can be caused by something else
  - **Smooth I(q):** particular protein shape
  - **Porod exponent < 4**: particle shape, poor fitting
  - **Plateau or increase in dimensionless Kratky plot**: shape of macromolecule (e.g. more anisotropic), poor background subtraction
  - **No plateau in Porod-Debye plot**: shape of macromolecule
  - **Smooth P(r)**: particular protein shape
  - **Extended tail on P(r), D_{max} hard to determine**: aggregation
  - **Rg, I(0) from P(r) larger than Guinier**: poor determination of D_{max}
  - **Overestimates of M.W.**: Wrong oligomer or aggregate in solution, inherent uncertainty in M.W. methods (usually ~10%)
  - **Narrow Guinier range**: aggregation/repulsion
  - **Flory exponent**: poor fitting, bad choice of model
  - **Extended reconstructions or rigid body models**: Poor reconstructions (e.g. with high ambiguity), actual extended particle shape
- Should always combine several different indicators to make statement about flexibility
  - Determining flexibility from SAXS requires preponderance of evidence
  - Eliminate other possibilities, such as using P(r) shape to rule our extended rigid shapes vs. flexibility in normalized Kratky plot, good Guinier to discount aggregation



Analyzing flexible systems

- Flexible systems do not exist in a single conformation in solution
- Analysis can address what range of conformations are in solution, in some cases what an average conformation may look like
- Ensemble methods for analysis describe set of possible conformations, set of actual conformations in solution
- Single conformation methods may describe an average conformation, should only be used for tightly peaked ensembles
  - Bead models incorporate no flexibility, can be useful for determining relative position of folded and flexible regions (Receveur-Brechot and Durand, 2012. DOI: 10.2174/1389203127992779O1)
  - CORAL, BUNCH (ATSAS) fit flexible linkers to known folded domains/subunits, usually result in a model that is close to size of the average conformation in the ensemble (Bernado, 2010. DOI: 10.1007/s00249-009-0549-3)
  - Cannot describe the full range of conformations, so proceed with caution or not at all!



Ensemble analysis

- Flexible systems sample a large number of conformations, so ensembles of conformations are the most appropriate way to represent the state in solution

- Ensemble analysis methods use the following approach:
  1. Computational generation of a large ensemble describing the conformational landscape available to the protein
  2. Computation of theoretical SAXS profiles from the individual conformations
  3. Selection of a sub-ensemble of conformations that collective describes the experimental profile

  (Cordeiro et al., 2017. DOI: 10.1007/978-981-10-6038-0_7)

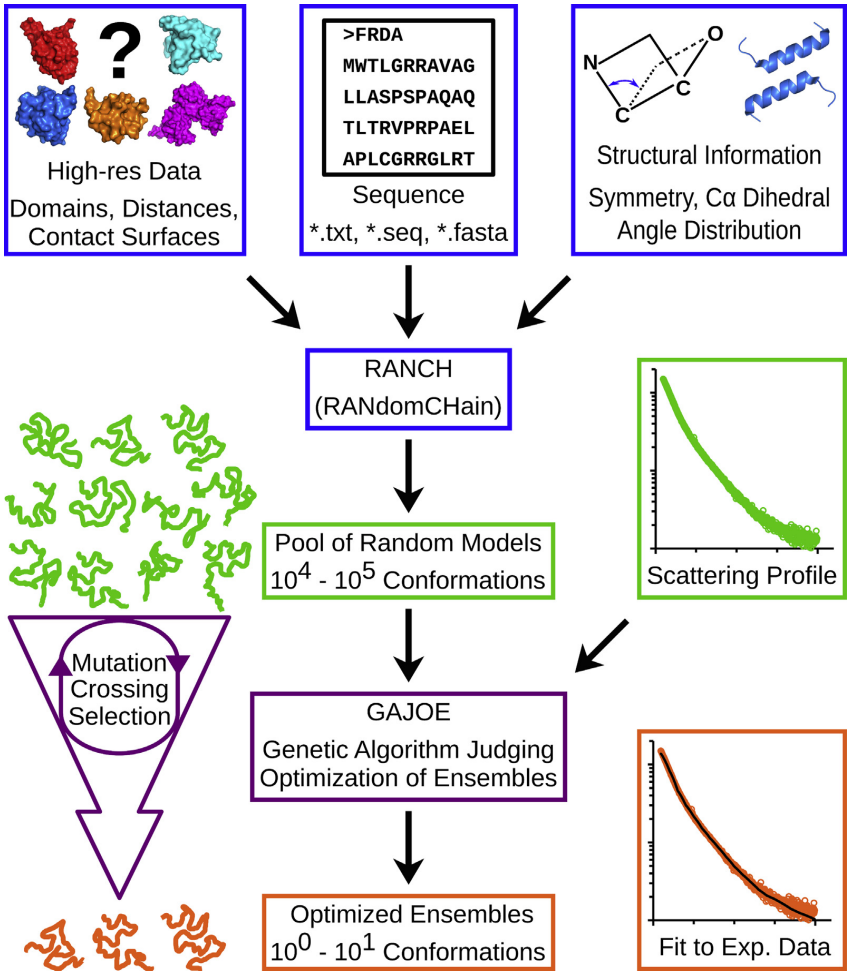
- Several different programs exist for ensemble analysis:
  - ASTEROIDS
  - Basis-Set Supported SAXS (BSS-SAXS)
  - Bayesian Ensemble SAXS (BE-SAXS)
  - Broad Ensemble Generator with Re-weighting (BEGR)
  - Ensemble Optimization Method (EOM, ATSAS)
  - Ensemble Refinement of SAXS (EROS)
  - ENSEMBLE
  - Minimal Ensemble Search (BilboMD-MES)
  - Maximum Occurrence (MAX-Occ)
  - MultiFoXS
  - SASSIE
  - Probably more . . .

- Programs differ in how main ensemble pool is generated, how profiles are calculated, how sub-ensemble is selected

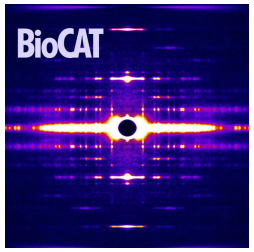
EOM

- Ensemble Optimization Method (EOM) from the ATSAS package was the original approach, remains the most widely used method

Tria et al., 2015.
DOI: 10.1107/S205225251500202X

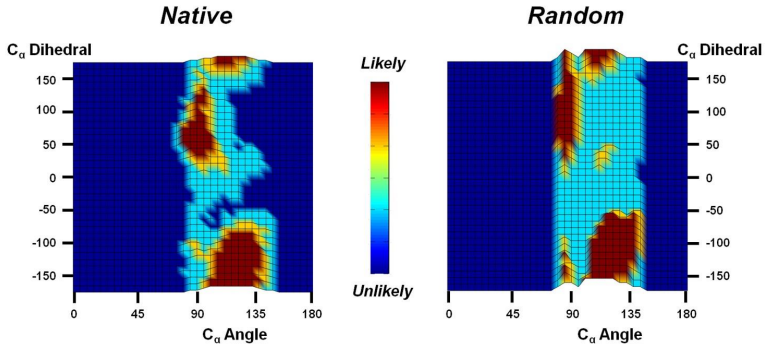


Grawert and Svergun, 2020.
DOI: 10.1016/j.jmb.2020.01.030

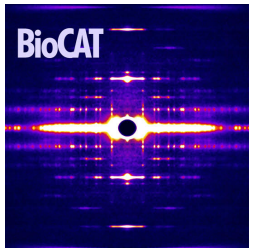


EOM – Generating a pool of structures with RANCH

- Pool generation requires:
  - Realistic and adequate sampling of conformational space
  - Produce feasible models (e.g. avoid steric clashes)
  - Incorporate high resolution information if available
  - Without a good starting pool, EOM results will not be valid
- Uses an algorithm based on bond vs. dihedral angle distribution represented by Cα-Cα Ramachandran plot
  - Allows generation of models resembling chemically denatured or natively unfolded proteins
- Can incorporate high resolution information for multiple domains, with either fixed or free positions
- Oligomers can be generated via symmetry operations

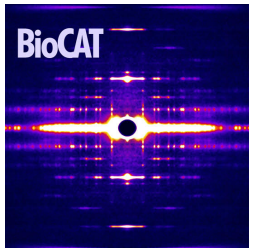


Tria et al., 2015.
DOI: 10.1107/S205225251500202X



EOM – Generating a pool of structures with RANCH

- Generates flexible structures from a single sequence
  - No multi-chain proteins unless generatable through symmetry operators
- If protein consists of known folded domains, can input one or more high resolution structures (.pdb)
- Input structures are kept fixed, any amino acids not in the high resolution structures are allowed to move
  - So delete any flexible loops/linkers from input structures if they have been modeled in
- Input high resolution structures and input sequences must match exactly, amino acid by amino acid
  - If there are gaps in your input .pdb file (e.g. missing flexible loop) you must split the structure into multiple .pdb files around those gaps
  - Sequence should also exactly match the sample, including tags and post-translational modifications
- Can constrain symmetry, add known distance constraints/contacts
- Can pick between different structure types for pool generation:
  - Random-coil (default) - CA dihedral angle distribution consistent with chemically denatured proteins
  - **Native-like (recommended) - CA angle distribution consistent with disordered proteins**
  - Compact-chain – CA angle distribution consistent with disordered proteins, forces more compact linkers
- Default is to generate 10,000 models, I prefer 50,000
  - Do a test run first with ~100 models to make sure settings are correct. Save the generated models and open a few in PyMOL or similar to see if reasonable conformations are being generated

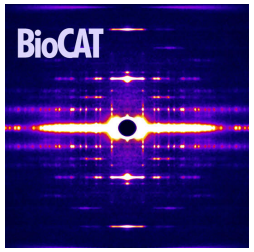


EOM – Selecting a sub-ensemble with GAJOE

- A genetic algorithm is employed to select a sub-ensemble
  1. Select 50 (default) sub-ensembles of structures
     - Number of structures in an ensemble can be dynamically selected by program. Usually 5-20.
  2. Generate 100 new sub-ensembles by:
     1. Replacing 20% of structures in each sub-ensemble with structures from pool or other sub-ensemble
     2. Exchanging at least 2 structures between two randomly chosen sub-ensembles
  3. Calculate fit for each of the 150 sub-ensembles against data
  4. Select the 50 best fit sub-ensembles
  5. Repeat the steps 2-4 1000 (default) times and take the final best fitting sub-ensemble

- The genetic algorithm is run multiple times to generate a final sub-ensemble:
  - Repeat the entire sub-ensemble selection process, steps 1-5, 100 (default) times
  - Combine all the best sub-ensembles from all 100 runs into a single final sub-ensemble
  - Calculate the scattering profile, R_g, D_{max}, Volume, and average Cα-Cα distance distributions for this final sub-ensemble. This represents the conformational space accessed by the protein in solution

- The genetic algorithm also outputs the structures in the single best fit sub-ensemble
  - One the small ~5-20 member sub-ensembles output from a single refinement of the genetic algorithm, steps 1-5
  - Does not represent every possible state in solution, just a few representative conformations. Do not over interpret these structures!



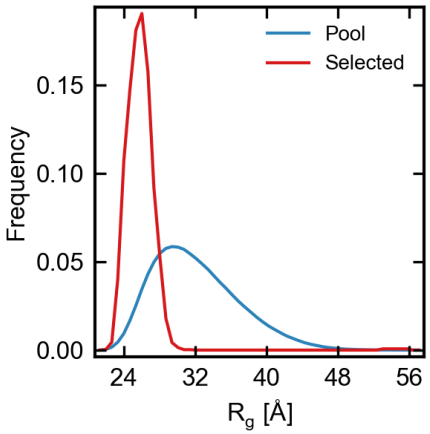
EOM – Selecting a sub-ensemble with GAJOE

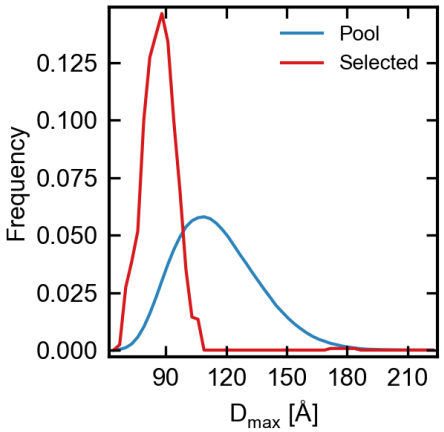
- GAJOE can use a pool of structures generated by RANCH (typical) or by another program
- Multiple pools of structures can be used
  - E.g. combining monomer and dimer pools
  - Generally not recommended
- Make sure to specify a sufficient number of spherical harmonics for calculating the scattering profile of each structure
  - 15 (default) fine for compact structures. Large particles, like IDPs, should use more
  - Can test by generating a few structures, calculating their profiles with CRYSOL with different numbers of harmonics and seeing if the profile changes
- I typically run GAJOE 10 independent times, compare the results to make sure the algorithm has sufficiently converged
- EOM can be run as a single command (eom), or RANCH and GAJOE can be run separately.
  - GAJOE can be re-run on the same pool without rerunning RANCH

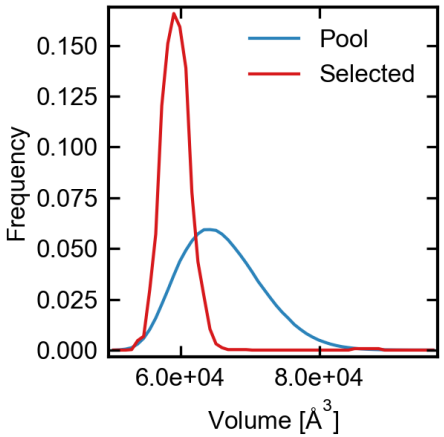
EOM - Results

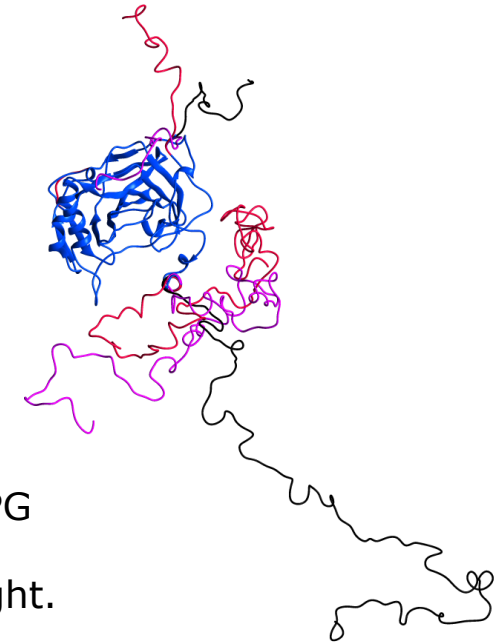
- Main EOM output is the R_g, D_{max}, and Volume distribution of the selected ensemble

- Comparison to the distributions from the full pool of conformational space allows you to make statements about degree of flexibility of the protein







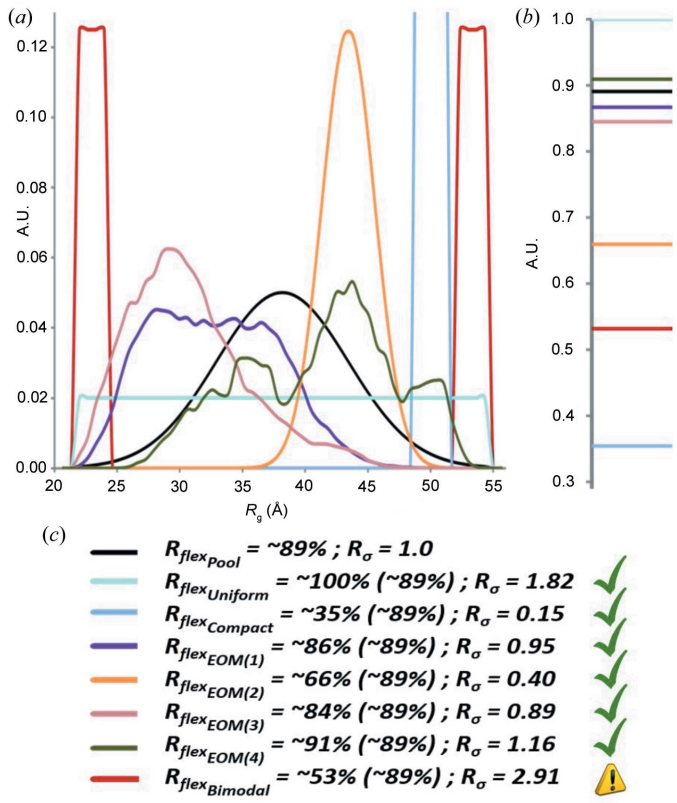


Comparison of selected ensemble vs. pool values shows that CA IX's PG domain (an IDR) adopts primarily compact conformations close to the folded catalytic domain. Sample ensemble structures shown on the right.

Koruza et al. 2020, DOI: 10.3390/ijms21155277

EOM - Results

- EOM also provides some advanced metrics of flexibility
- R_{flex} is a quantification of the flexibility of a pool (either total or selected ensemble)
  - Allows quantification between flexible and rigid systems
- R_σ is the ratio of the standard deviation of the selected ensemble and pool
  - Close to 1 when selected ensemble describes a fully flexible system and reproduces the conformational space of the pool
  - If the R_{flex} of the selected ensemble is smaller than that of the pool, we expect $R_\sigma < 1$
  - If the R_{flex} of the selected ensemble is larger than that of the pool, we expect $R_\sigma > 1$
  - If the R_{flex} of the selected ensemble is significantly smaller than that of the pool, and $R_\sigma > 1$, may be a problem with the EOM result
- There is currently a bug in the EOM calculation of R_σ, it is significantly off (e.g. 4.5 vs. 1.1 for a recent EOM run I did). Until ATSAS 3.0.3 is released R_σ should be recalculated directly from the output EOM distributions!

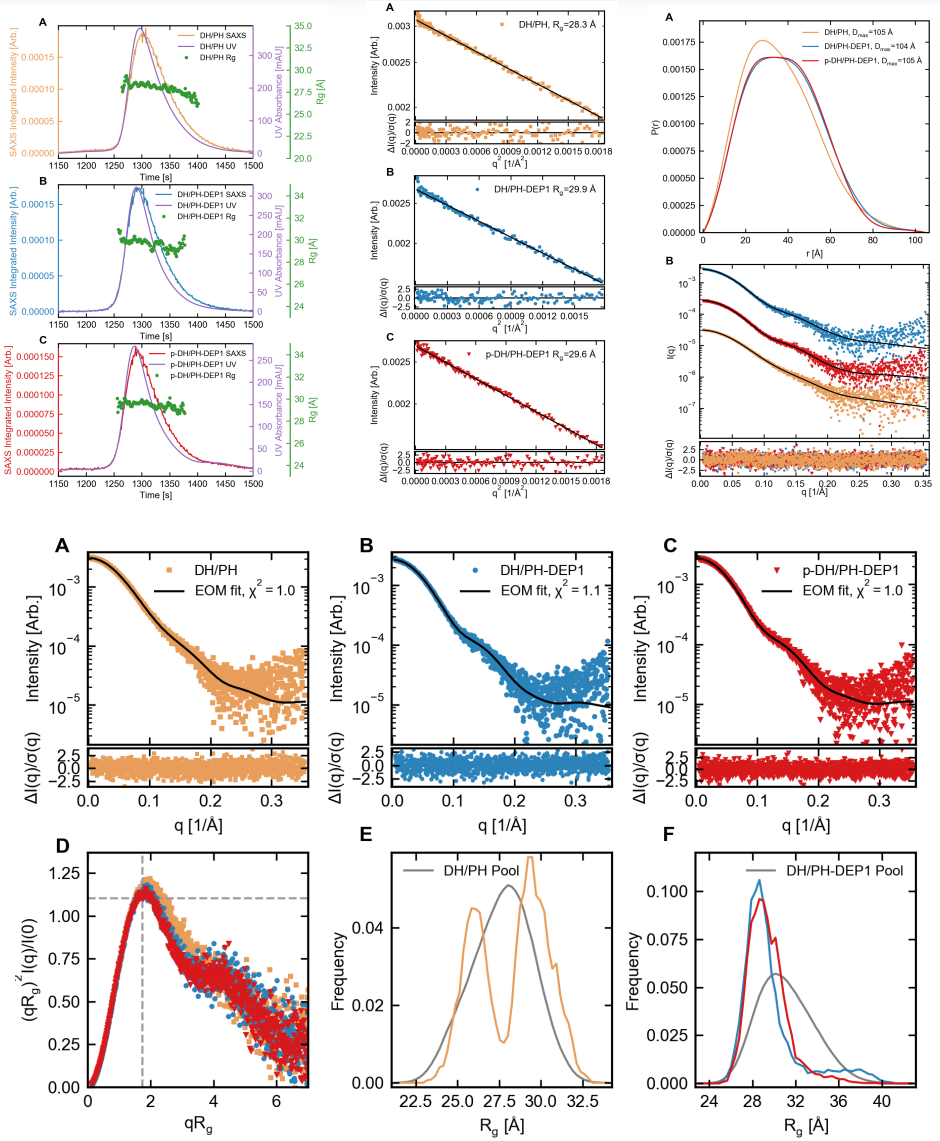


Different distributions, and their R_{flex} and R_σ values. Pool represents complete randomness.

Tria et al., 2015.
DOI: 10.1107/S205225251500202X

EOM – Example 1

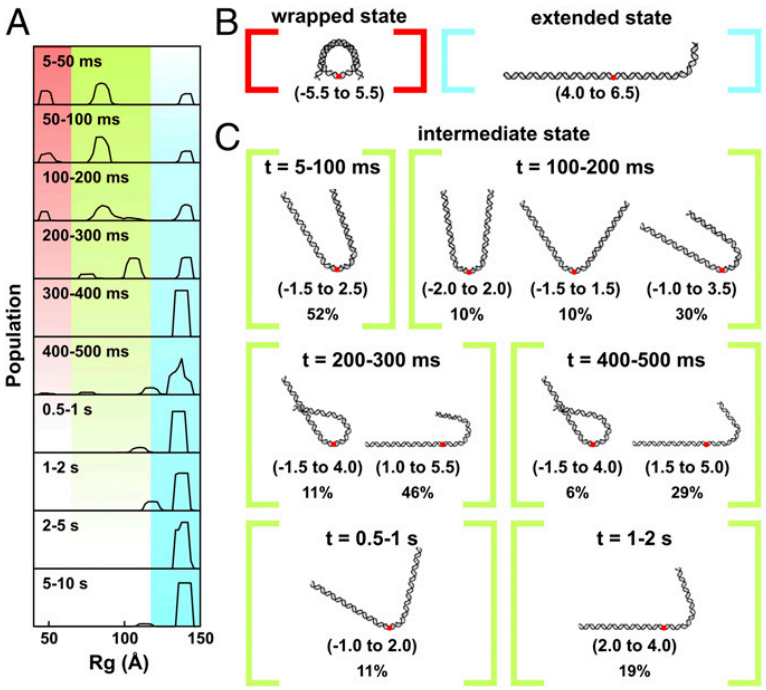
- Tesmer lab (Purdue) studies P-Rex1, a possible therapeutic target for cancer, neurological disorders, inflammatory diseases, and type 2 diabetes
- Studying C terminal fragments DH/PH and DH/PH-DEP1 domain constructs as DEP1 may be autoinhibitory.
- Significant flexibility in linker between DH/PH and DEP1
- SAXS combined with other structural and biophysical techniques was able to show that the DH/PH-DEP1 construct adopts significantly compact shapes in solution, most likely attributable to interaction between DH/PH and DEP1
- Based on this association, proposed a mechanism for inhibition



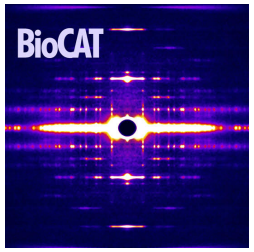
Ravala et al., 2020. DOI: 10.1074/jbc.RA120.014534

EOM – Example 2

- Pollack lab (Cornell) studies nucleosome dynamics
- Interested in DNA unwinding from the histone core, essential for transcription, replication, and repair
- Carried out time resolved contrast matching SAXS to study dynamics of unwinding
- Each measured timepoint contains a continuous distribution of different unwinding states, had to be fit with an ensemble
- Generated a pool of candidate structures using custom methods, carried out ensemble selection using GAJOE at each time point
- Additional SAXS data from equilibrium conditions at different unwinding stages, FRET data used to verify results
- Put together kinetic picture of DNA states during unwinding

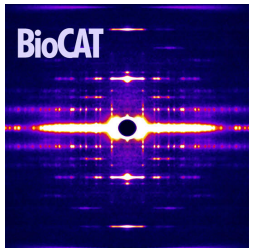


Chen and Tokuda et al., 2017. DOI: 10.1073/pnas.1611118114



Summary

- SAXS is a powerful tool for studying flexible systems in solution
  - One of the few methods that can quantitatively characterize partially disordered or completely disordered macromolecules
- Use multiple indicators to determine whether a system is flexible
  - Other effects, such as aggregation or extended overall shape can show same effects on profile and parameters as flexibility
- Use ensemble analysis to inform about overall set of conformations sampled in solution by your system
  - Rarely if ever is it appropriate to discuss a single state of the flexible system in solution
- EOM in the ATSAS package is the most common tool, but many others are available



References

- Overviews:
  - Brosey, C. A. & Tainer, J. A. (2019). *Curr. Opin. Struct. Biol.* **58**, 197–213.
  - Kikhney, A. G. & Svergun, D. I. (2015). *FEBS Lett.* **589**, 2570–2577.
  - Receveur-Brechot, V. & Durand, D. (2012). *Curr. Protein Pept. Sci.* **13**, 55–75.
  - Bernadó, P. & Svergun, D. I. (2012). *Methods in Molecular Biology*, Vol. *896*, pp. 107–122.
  - Rambo, R. P. & Tainer, J. A. (2011). *Biopolymers*. **95**, 559–571.
  - Bernadó, P. (2010). *Eur. Biophys. J.* **39**, 769–780.

- EOM:
  - Bernado, P., Mylonas, E., Petoukhov, M. V, Blackledge, M. & Svergun, D. I. (2007). *J. Am. Chem. Soc.* **129**, 5656–5664.
  - Tria, G., Mertens, H. D. T., Kachala, M. & Svergun, D. I. (2015). *IUCrJ*. **2**, 207–217.