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
Monocromatic X-ray beam

Sample
Mono- or polydisperse

\[ I(s) = \sum_N I_j(s) = \sum_K \nu_k I_k(s) \]

\[ s = \frac{4\pi \sin \theta}{\lambda} \]

Data analysis with ATSAS

\[ I(s) = \langle I(s) \rangle_\Omega = (A(s) \cdot A^*(s))_\Omega \]
Small Angle X-ray Scattering

Monocromatic X-ray beam
Monocromatic
Sample
Mono- or polydisperse

\[ I(s) = \sum_{N} I_{j}(s) = \sum_{K} \nu_{k} I_{k}(s) \]

Methods to study IDPs:

<table>
<thead>
<tr>
<th>Method</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMR</td>
<td>High resolution&lt;br&gt;Limited size of the protein</td>
</tr>
<tr>
<td>SAXS</td>
<td>Overall parameters&lt;br&gt;Fast measurement – a lot of various conditions</td>
</tr>
<tr>
<td>FRET</td>
<td>Single-molecule approach&lt;br&gt;Protein must be modified</td>
</tr>
</tbody>
</table>

Data analysis with ATSAS

\[ I(s) = \langle I(s) \rangle_{\Omega} = \langle A(s) \cdot A^*(s) \rangle_{\Omega} \]
Can I use this data for further analysis?

Initial Data Analysis: Data Quality

“Can I use this data for further analysis?”

Log I(s)

Log I(s)

AGGREGATED!
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 \( (I(s) \cdot s^2 \text{ vs } s) \)

   [Image: Kratky plot for three constructs of Src-Kinase]
   - Folded
   - Unfolded
   - Both Folded and Unfolded

   \[ \frac{I(s)}{I(0) \cdot s^2} \text{ vs } s (\text{nm}^{-1}) \]

2. Radius of gyration is a single parameter


*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

**BAD FIT**
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

GOOD FIT!
Selected ensemble

Pool

EOM: Rg Distribution
Examples of EOM application


And many more...
Questions to Answer

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

Number of chains: 10

Number of chains: 100

Number of chains: 1 000

Number of chains: 5 000

Number of chains: 10 000

Number of chains: 64 790

Globular protein (Rg = 13.52)

100-AA-alpha-helix (Rg = 35.95)
EOM Tests: Resolving Open and Closed Conformations

<table>
<thead>
<tr>
<th>Proteins</th>
<th>$R_g$ open conformation (CRYSOL)</th>
<th>$R_g$ closed conformation (CRYSOL)</th>
<th>Min $R_g$ of the pool</th>
<th>Max $R_g$ of the pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenylate kinase</td>
<td>20.88</td>
<td>17.98</td>
<td>24.82</td>
<td>75.92</td>
</tr>
<tr>
<td>Calmodulin</td>
<td>21.53</td>
<td>16.29</td>
<td>20.84</td>
<td>66.56</td>
</tr>
<tr>
<td>Tryptophan repressor</td>
<td>30.61</td>
<td>18.59</td>
<td>18.40</td>
<td>58.38</td>
</tr>
</tbody>
</table>
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

100 AA

15-20% Undistinguishable
$\Delta R_g = 11.33$

10-15% Well distinguishable
$\Delta R_g = 13.02$
Absolute Resolution ($\Delta$) is the minimal difference in average $R_g$ between two populations still distinguishable on $R_g$ distribution.

Relative Resolution ($\delta$) is ratio between Resolution and standard deviation of the pool:

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

If $\delta > 2.3$ then EOM can be used to distinguish two populations.
Quantitative Parameters

Absolute values

- Radius of gyration ($R_g$)
- Width of GAJOE distribution ($\sigma$)

Relative parameters

- $a = f(R_g, N)$
- $b = \sigma_{EOM}/\sigma_{pool}$
Quantitative parameters: Determination of $a$ and $b$

\[
R_g = 1.927 \cdot N^{0.598}
\]

$\alpha = 1$

\[
R_g = 2.83 \cdot N^{0.34}
\]

$\alpha = 0$
Quantitative parameters: Determination of $a$ and $b$

$R_g = 1.927 \cdot N^{0.598}$

$R_g = 2.83 \cdot N^{0.34}$

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

Flexible as the pool $b = 1$

Rigidity border is $b = 0.64$
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

Log I(s)

Resolution, nm

2.00 1.00 0.67 0.50 0.33

Size

s, nm⁻¹

Atomic structure

Fold

Shape

© Dmitri Svergun
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** ($\Delta$) is the minimal difference in average $R_g$ between two populations still distinguishable on GAJOE Rg distribution
- **Relative Resolution** ($\delta$) is ratio between Resolution and standard deviation of the pool:
  \[ \delta = \frac{\Delta}{\sigma} \]

**Estimation of the resolution:**

<table>
<thead>
<tr>
<th>Case</th>
<th>$\sigma$, Å</th>
<th>Smallest distinguishable $\Delta R_g$, Å</th>
<th>$\delta$ estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>100AA, wide</td>
<td>5.52</td>
<td>10.86 $&lt; \Delta &lt; 12.85$</td>
<td>1.97 $&lt; \delta &lt; 2.32$</td>
</tr>
<tr>
<td>100AA, narrow</td>
<td>5.52</td>
<td>11.33 $&lt; \Delta &lt; 13.02$</td>
<td>2.05 $&lt; \delta &lt; 2.36$</td>
</tr>
<tr>
<td>200AA, narrow</td>
<td>8.52</td>
<td>16.21 $&lt; \Delta &lt; 19.90$</td>
<td>1.90 $&lt; \delta &lt; 2.34$</td>
</tr>
</tbody>
</table>

If $\delta$ is more than **2.3** then GAJOE can be used to distinguish two populations
Quantitative Parameters: Real-life Cases

\[ b = \frac{\sigma_{GAJOE}}{\sigma_{pool}} \]

\[ a = \frac{R_g - R_g^{Globular}}{R_g^{Random~Coil} - R_g^{Globular}} \]
Initial Data Analysis: Background Subtraction

**Solution minus Solvent**

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

Random Pool

Genetic Algorithm

Deconvolution

Random Selection

C Chromosomes

N Conformations

Mutations

Crossing

G Generations

R EOM Runs

Comparison of Distributions:
Quantitative structural information